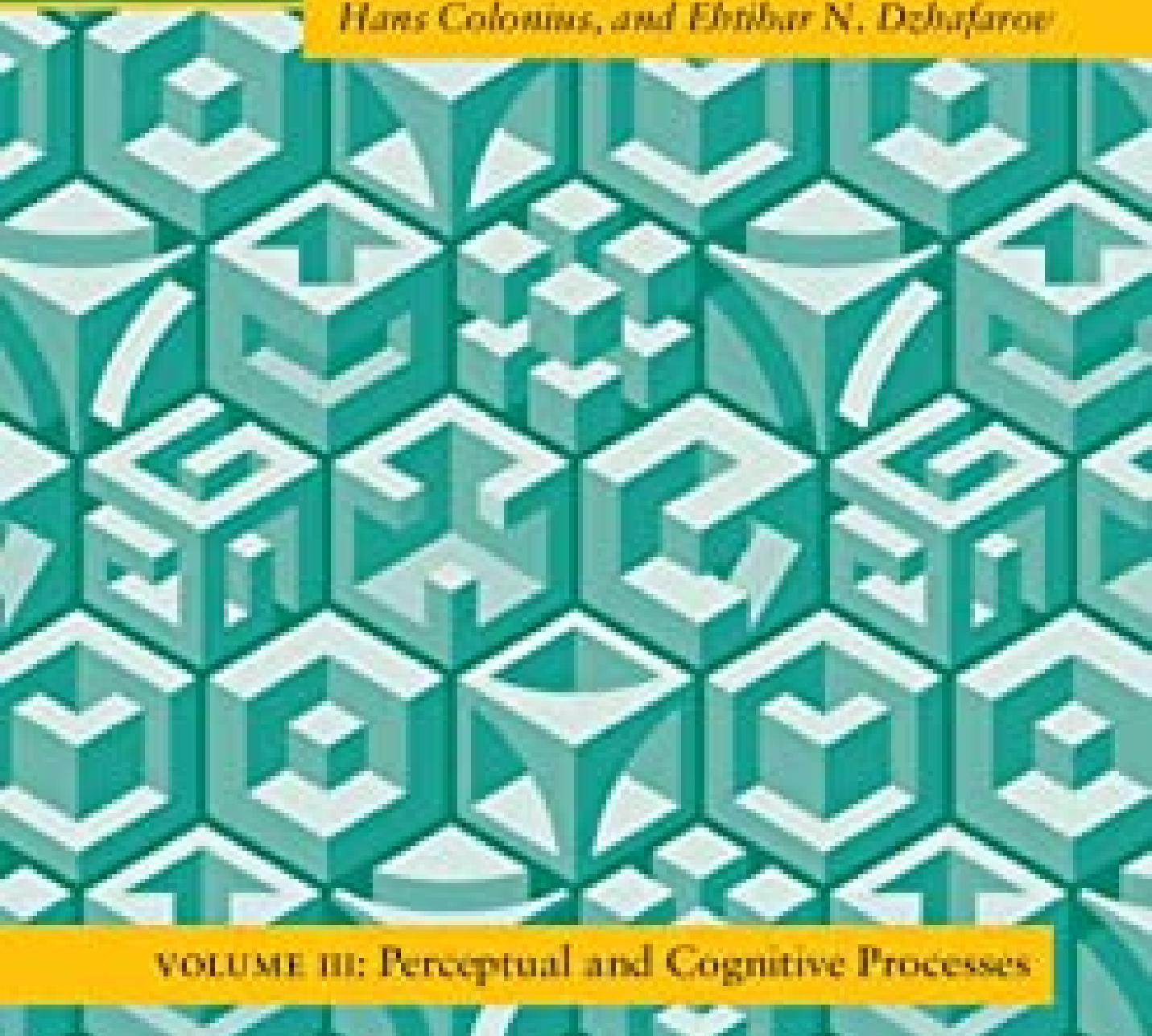
NEW HANDBOOK OF MATHEMATICAL PSYCHOLOGY

Edited by F. Gregory Ashby, Hans Colonius, and Ehtibar N. Dzhafarov



VOLUME III: Perceptual and Cognitive Processes

New Handbook of Mathematical Psychology

Volume 3: Perceptual and Cognitive Processes

The field of mathematical psychology began in the 1950s and includes both psychological theorizing, in which mathematics plays a key role, and applied mathematics motivated by substantive problems in psychology. Central to its success was the publication of the first *Handbook of Mathematical Psychology* in the 1960s. The psychological sciences have since expanded to include new areas of research, and significant advances have been made both in traditional psychological domains and in the applications of the computational sciences to psychology. Upholding the rigor of the original Handbook, the *New Handbook of Mathematical Psychology* reflects the current state of the field by exploring the mathematical and computational foundations of new developments over the last half-century. The third volume provides up-to-date, foundational chapters on early vision, psychophysics and scaling, multisensory integration, learning and memory, cognitive control, approximate Bayesian computation, and encoding models in neuroimaging.

F. GREGORY ASHBY is Distinguished Professor Emeritus of Psychological and Brain Sciences at the University of California, Santa Barbara, USA. He is the author of more than 180 publications, including nine articles in *Psychological Review* and four books. He is past president of the Society for Mathematical Psychology and past chair of the NIH Cognition and Perception Study Section. His awards include the Howard Crosby Warren Medal in 2017.

HANS COLONIUS is Professor of Psychology at Oldenburg University, Germany. He has published 130 papers and two books. He was editor-in-chief of the *Journal of Mathematical Psychology* and a visiting professor at various universities across Europe and the USA. His awards include a Heisenberg professorship in 1982, and his work has been supported by numerous grants from the German Science Foundation.

EHTIBAR N. DZHAFAROV is Professor of Psychological Sciences at Purdue University, USA. He has published 170 papers in psychology, mathematics, philosophy, and foundations of quantum mechanics; he has also edited six books and three special journal issues. He served as president of the Society for Mathematical Psychology, and he has received a Humboldt Research Award.

New Handbook of Mathematical Psychology

Volume 3. Perceptual and Cognitive Processes

Edited by

F. Gregory Ashby
University of California, Santa Barbara

Hans Colonius
Carl V. Ossietzky Universität Oldenburg, Germany

Ehtibar N. Dzhafarov
Purdue University, Indiana

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108830676

DOI: 10.1017/9781108902724

First published 2023

A catalogue record for this publication is available from the British Library.

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-108-83067-6 Hardback

Contents

Contributors

F. GREGORY ASHBY, University of California, Santa Barbara (USA)

DAVID BRAINARD, University of Pennsylvania (USA)

HANS COLONIUS, Oldenburg University (Germany)

MATTHEW J. CROSSLEY, Macquarie University (Australia)

JIMMY DE LA TORRE, The University of Hong Kong (Hong Kong)

ADELE DIEDERICH, Oldenburg University (Germany)

EHTIBAR N. DZHAFAROV, Purdue University (USA)

MARC W. HOWARD, Boston University (USA)

JEFFREY B. INGLIS, University of California, Santa Barbara (USA)

MIGUEL A. SORREL, Autonomous University of Madrid (Spain)

FABIÁN A. SOTO, Florida International University (USA)

NOAH THOMAS, The Ohio State University (USA)

BRANDON M. TURNER, The Ohio State University (USA)

TRISHA VAN ZANDT, The Ohio State University (USA)

BRIAN WANDELL, Stanford University (USA)

MICHAEL J. WENGER, University of Oklahoma (USA)

Preface

In 1845 Edgar Allan Poe published a story titled "The Purloined Letter," in which a protagonist, Mr. C. Auguste Dupin, says the following:

> The mathematics are the science of form and quantity; mathematical reasoning is merely logic applied to observation upon form and quantity. The great error lies in supposing that even the truths of what is called pure algebra, are abstract or general truths. And this error is so egregious that I am confounded at the universality with which it has been received. Mathematical axioms are not axioms of general truth. What is true of relation — of form and quantity — is often grossly false in regard to morals, for example. In this latter science it is very usually untrue that the aggregated parts are equal to the whole. [...] two motives, each of a given value, have not, necessarily, a value when united, equal to the sum of their values apart. There are numerous other mathematical truths which are only truths within the limits of relation. But the mathematician argues, from his finite truths, through habit, as if they were of an absolutely general applicability — as the world indeed imagines them to be.

A safe reaction to this excerpt (especially in view of Mr. Dupin's subsequent remarks, omitted here) is that Mr. Dupin has a hopelessly approximate notion of mathematics. However, his appellation to morals and motives provides an opportunity for a more generous reaction, making Mr. Dupin's tirade relevant to a discussion of mathematical psychology. One could interpret this tirade as stating that

D1 given two motives or moral ideas A and B that are combined in some well-defined sense (e.g., co-occur chronologically),
D2 and assuming that each of them can be assigned a value represented by a real number, a and b,
D3 and assuming that the combination of A and B can also be assigned a value c that is a real number,
D4 and assuming that the combination of A and B is represented by the sum of their individual values, $a + b$,
D5 we observe empirically that the value c is not generally equal to $a + b$;
D6 the contradiction between D4 and D5 shows that the laws of arithmetic do not apply to motives and moral ideas.

Of course, the assumptions D1–D4 are hidden, they are not explicated by Mr. Dupin. Nor would he stop to think about how he could know the truth of D5. Deny any of the assumptions D1–D5, and Mr. Dupin will lose any grounds to blame mathematics. For instance, if the assumption D4 is not made, then c does not have to be equal to $a + b$, it can instead be ab or $\max(a, b)$, or perhaps a and b alone do not determine c at all. Mathematics is perfectly fine with these possibilities. Mathematics is also fine with the possibility that the assumptions D2 and D3 are wrong, and the motives or moral ideas are not representable by anything that can be subjected to conventional addition. Perhaps a and b are dimensioned numbers, but their dimensionality is not the same (say, they are measured in "love units" and "revenge units," respectively).

Is there any useful lesson that can be derived from this admittedly too easy critique of Mr. Dupin's perorations? We think there is. The lesson is that mathematics in psychology (or chemistry, or wherever else it is applied) is not about adding, multiplying, or, generally, computing. It is primarily about striving for conceptual clarity and avoiding conceptual confusions. Before we can compute, we need to explicate the hidden assumptions we make, and often when we do this we find out these assumptions are not all that compelling.

Take as an example the following piece of reasoning one can encounter in the modern literature. In logic, the conjunction of two statements is commutative, $A\&B$ is the same as $B\&A$. However, we have empirical evidence that the chronological order in which two statements are presented or evaluated does matter for one's judgment of the truth value (or probability) of their conjunction. Ergo, classical logic (probability theory) is not applicable to human judgments. Let us see what is involved in this reasoning.

L1 Assuming that if A is presented first and B is presented second, then their combination is represented by $A\&B$,
L2 whence, by symmetry, if A is presented second and B is presented first, their combination is represented by $B\&A$;
L3 and knowing from classical logic that $A\&B$ and $B\&A$ are equivalent,
L4 their truth value (or probability) M should be the same, $M(A\&B) = M(B\&A)$.
L5 But empirical observations tell us this is not generally the case.
L6 Ergo, classical logic (classical probability) here does not work.

The reasoning here is definitely "Dupinesque." Far from not being applicable, formal logic, if applied correctly, would lead one to reject, by *reductio ad absurdum*, the assumptions L1 and L2. Indeed, L3 and the implication L3→L4 are unassailable, and we assume L5 truthfully describes empirical facts. The ways to constructively deny L1 and L2 readily suggest themselves. One way is to introduce a special, noncommutative operation *A then B*. Another way is to identify the statements not only by their content but also by their chronological position in the pair: a statement with content A, if presented first, is A_1, when presented second it is A_2. So the rejected representations $A\&B$ and $B\&A$ in L1 and L2 are in reality $A_1\&B_2$ and $A_2\&B_1$, respectively. The commutativity of the conjunction is

perfectly preserved, e.g., $A_1\&B_2 \equiv B_2\&A_1$. But $A_1\&B_2$ and $A_2\&B_1$ are different propositions, and one should generally expect that

$$M(A_1\&B_2) \neq M(A_2\&B_1),$$

whatever M may be. One can further investigate which of the two solutions, the introduction of *A then B* or the positional labeling, is preferable. Thus, if the truth values of the statements A and B themselves, and not just of their conjunction, depends on their chronological position, then the positional labeling clearly wins.

The quest for conceptual clarity and explication of hidden assumptions often faces greater and subtler difficulties than in the examples above. The greater then are the rewards ensuing from resolving these difficulties. Take as an illustration the question of whether the ways we measure certain quantities constrain the way these quantities can be related to each other. The historical context for this question is the emergence in mathematical psychology in the second half of the twentieth century of the line of research referred to as representational theory of measurement. It is an unusual theory, in the sense that while it is a firmly established branch or part of mathematical psychology, its aim is to formalize all empirical measurements, across sciences, and even provide necessary conditions for all possible scientific laws and regularities.

One of the tenets of this theory, widely accepted in modern psychology (and in textbooks of elementary statistics), is that all entities we deal with, physical or mental, are measured on specific scales, such as ordinal, interval, or ratio scales. We need not get here into the details of the qualitative, or pre-numerical, symmetries (automorphisms) postulated for the entities being measured. Suffice it to mention that the scale type assigned to these entities is characterized by the class (usually, a parametric group) of interchangeable mathematical representations, i.e., measurement functions, mapping the entities being measured into mathematical objects, usually real numbers. Thus, if entities $\mathbf{x} \in \mathcal{X}$, say, stimulus intensity or sensation magnitude values, are said to be measured on a ratio scale, it means that the measurement functions for $\mathcal{X}$ map this set into intervals of real numbers, and that if f and g are such measurement functions, then, for every $\mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{x}) = kg(\mathbf{x}),$$

for some positive constant k. R. Duncan Luce, arguably one of the two greatest mathematical psychologists of the twentieth century (along with William K. Estes), made use of the notion of a measurement scale to restrict theoretically the class of possible psychophysical functions, those relating the magnitude of stimulus to the magnitude of sensation it causes. Luce proposed this idea in a book entitled *Individual choice behavior* (Luce, 1959a) and in a journal article (Luce, 1959b). The idea is so attractive aesthetically that it deserves being reproduced here, *mutatis mutandis*.

Let $x = f(\mathbf{x})$ and $s = \varphi(\mathbf{s})$ represent measurement functions for the stimulus magnitude $\mathbf{x}$ and sensation magnitude $\mathbf{s}$, respectively, and let the

psychophysical function relating **s** to **x**, written in terms of these specific measurement functions, be

$$s = \psi(x).$$

Assume that both **x** and **s** are of the ratio-scale type. Consider another admissible measurement function for **x**:

$$x' = kf(\mathbf{x}),$$

for some $k > 0$. Then, Luce hypothesized, if one switches from x to x', the psychophysical function should be presentable as

$$s' = \psi\left(x'\right),$$

where

$$s' = c\varphi(\mathbf{s}),$$

for some $c > 0$. That is, s' is another admissible measurement function for **s**. Put differently, the function ψ is invariant with respect to admissible changes of the measurement function for **x**, provided that the measurement function for the dependent variable **s** can also change to other measurement functions accordingly. The last word, "accordingly," means that the choice of the measurement function for **s** generally depends on the choice of the measurement function for **x**, i.e.,

$$c = K(k),$$

for some function K.

The reasoning here is seductively plausible, and Luce thought that examples of the well-established laws of physics confirmed its validity. Thus, Newton's law of gravitation is conventionally written as

$$F = \gamma \frac{m_1 m_2}{r^2}.$$

If we assume that everything on the right-hand side is fixed except for the distance measurement function r, then augmenting this measurement function by the factor of $k = 10$ would result in the same expression, except that the measurement function F for force will have to be multiplied by $c = k^{-2} = 1/100$.

Having accepted Luce's hypothesis (Luce called it a "principle of theory construction"), we are led to a surprising conclusion: the psychophysical function cannot be anything but a power function. What is surprising here is that this conclusion is based on no empirical evidence, it is obtained deductively, by merely assuming that the magnitudes of stimulus and sensation are of the ratio-scale type. Indeed, the reasoning above translates into

$$\psi(kx) = \psi\left(x'\right) = s' = cs = c\psi(x) = K(k)\,\psi(x),$$

whence, by eliminating all but the marginal terms, we get the functional equation

$$\psi(kx) = K(k)\,\psi(x).$$

Here, the values of x and k are positive, and the functions ψ and K are positive and increasing. Since the functional equation holds for all positive k and all x on some interval of positive reals, its only solution is known to be (Aczel, 1987)

$$\psi(x) = bx^{\beta}, K(k) = k^{\beta},$$

for some positive b and β.

It looks like we have here an immaculate piece of deductive reasoning, with all concepts rigorously defined and all assumptions explicated. However, what shall we do with the fact that psychophysical laws of other forms have been proposed too? Most notably, every psychologist knows of the logarithmic law proposed by Gustav Theodor Fechner in 1861:

$$s = s_0 \log \frac{x}{x_0}.$$

Here, x_0 is the numerical representation of the absolute threshold magnitude $\mathbf{x_0}$, one at which the numerical representation of $\mathbf{s}$ is zero, for all measurement functions.

We can see that Fechner's law does not violate any of Luce's assumptions. Since $\mathbf{x}$ and $\mathbf{x_0}$ are measured by the same measurement function, the value of

$$\frac{f(\mathbf{x})}{f(\mathbf{x_0})} = \frac{kx}{kx_0}$$

is the same for all admissible f. The magnitude of the absolute lower threshold is defined irrespective of the measurement function chosen for $\mathbf{x}$, because so is defined $s = 0$. Even if one denies the existence of absolute threshold as a fixed constant, such operational definitions of $\mathbf{x_0}$ as "the value of $\mathbf{x}$ detected with probability p" are independent of the measurement function for $\mathbf{x}$. The measurement function for the dependent variable $\mathbf{s}$ is chosen independently, which formally translates into $K(k) = 1$. The value s_0 is the numerical representation of the value of $\mathbf{s}$ corresponding to the value of $\mathbf{x}$ at which $\log \frac{x}{x_0} = 1$.

Since the logarithmic law is not the same as the power law, Luce must have made a hidden assumption that Fechner's derivation of his law violates. This hidden assumption is not difficult to detect. It is the assumption that the dependence of $\mathbf{s}$ on $\mathbf{x} \in \mathcal{X}$ contains no parameters (constants with respect to $\mathbf{x}$) that belong to the same set $\mathcal{X}$ and are therefore represented by the same measurement function. Such parameters are called measurement-dependent constants, or dimensional constants in the case of ratio scales. An expression

$$s = s_0 \psi\left(\frac{x}{x_0}\right),$$

with dimensional constants x_0 and s_0, can hold for any positive increasing function ψ. Using examples of physical laws, this was pointed out to Duncan Luce by William W. Rozeboom in a 1962 article (Rozeboom, 1962). Being a true scientist, Luce accepted this criticism and withdrew his "principle of theory construction" (Luce, 1962). Interestingly, in the formulation of this principle, Luce did in fact

mention dimensional constants: the form of the dependence ψ should be invariant, he wrote, "except for the numerical values of parameters that reflect the effect on the dependent variables of admissible transformations of the independent variables." This is precisely what dimensional constants are. Using Luce's own example of the universal gravitation law, in the formula

$$F = \gamma \frac{m_1 m_2}{r^2},$$

if one uses the distance–time–mass–force system of units, changing the dimensionality of mass or distance in no way leads to the change of the dimensionality of force. Rather, the dimensional constant γ, whose dimensionality is

$$force \cdot distance^2 \cdot mass^{-2},$$

changes its numerical value. In essence, γ is a coalesced form (using the expression coined by Percy Williams Bridgman) of the "individual" dimensional constants in the formula

$$\frac{F}{F_0} = \frac{\frac{m_1}{m_0}\frac{m_2}{m_0}}{\left(\frac{r}{r_0}\right)^2}.$$

The lesson we learn from the story of Duncan Luce's "principle of theory construction" is that hidden assumptions and lack of conceptual clarity due to the failure to explicate them can be present even in very rigorous treatments. Moreover, explication of these hidden assumptions, while resolving the issue at hand, leads to new conceptual problems and opens new avenues of conceptual research. In our example the new conceptual problems can be formulated thus:

P1 What is the nature of dimensional (more generally, measurement-dependent) constants in empirical laws? Where do they come from?

P2 How do we know the scale type (the group of admissible measurement functions) of a given entity? Is it imposed on the entity by the human mind, or is it objectively present in it, to be uncovered?

These questions are at the foundations of all empirical science, and it is an interesting historical fact that their development owes a great deal to mathematical psychology (see, e.g., Dzhafarov, 1995; Falmagne & Doble, 2018; Narens, 2007). This preface, of course, is not a place to discuss these questions in any detail.

About this Volume

This is the third, and concluding, volume of the *New Handbook of Mathematical Psychology*. In the same way as the first two volumes, it offers a representative sample of several branches of mathematical psychology. This volume focuses on sensory and perceptual processing, learning and memory, and cognition.

Chapter 1, written by Brian Wandell and David Brainard, surveys low-level encoding of visual information. Modern vision science is highly interdisciplinary,

combining ideas from physics, biology, and psychology. In recent years, deductive mathematics in vision science is often combined with computational modeling to add realism to the mathematical formulations. Together, the mathematics and computational tools provide a realistic estimate of the initial signals that the brain analyzes to render visual judgments of various aspects of visual image, such as motion, depth, and color. The chapter first traces the calculations from the representation of the light signal, to how that signal is transformed by the lens to the retinal image, and then how the image is converted into the cone photoreceptor excitations. The central steps in the initial encoding rely heavily on linear systems theory and the mathematics of signal-dependent noise. The chapter describes computational methods used to understand how light is encoded by cone excitations. The chapter also provides a mathematical formulation of the ideal observer that uses all the encoded information to perform a visual discrimination task, as well as Bayesian methods that combine prior information and sensory data to estimate the light input. These tools help one to reason about what information is present in the neural representation, what information is lost, and what types of neural circuits could extract information to make judgments about a visual scene.

Chapter 2, by Adele Diederich and Hans Colonius, deals with the topic of multisensory integration – that is, with the merging of the information provided by different sensory modalities. This topic has been the subject of many competing theories, often crossing boundaries between psychology and neuroscience. In defining the somewhat fuzzy term of "multisensory integration," it has been observed that at least some kind of numerical measurement assessing the strength of the crossmodal effects is always required. The focus of this chapter is on measures of multisensory integration based on both behavioral and single-neuron recording data: spike numbers, reaction time, frequency of correct or incorrect responses in detection, recognition, and discrimination tasks. On the empirical side, these measures typically serve to quantify effects on multisensory integration of attention, learning, and such factors as age, certain disorders, developmental conditions, training and rehabilitation. On the theoretical side, these measures often help to quantify important characteristics of multisensory integration, such as optimality in combining information or inverse effectiveness, without necessarily subscribing to any specific model of the mechanisms of multisensory integration.

Ehtibar Dzhafarov and Hans Colonius present a systematic theory of generalized (or universal) Fechnerian scaling in Chapter 3 that is based on the intuition underlying Fechner's original theory. This intuition is that subjective distances among stimuli are computed by means of cumulating small discriminability values between "neighboring" stimuli. A stimulus space is supposed to be endowed by a dissimilarity function, computed from a discrimination probability function for any pair of stimuli chosen in two distinct observation areas. On the most abstract level, one considers all possible chains of stimuli leading from a stimulus **a** to a stimulus **b** and back to **a**, and takes the infimum of the sums of the dissimilarities along these chains to be the subjective distance between **a** and **b**. In arc-connected spaces, the cumulation of dissimilarity values along all possible chains reduces to their

cumulation along continuous paths, leading one to a fully fledged metric geometry. In topologically Euclidean spaces, the cumulation along paths further reduces to integration along smooth paths, and the geometry in question acquires the form of a generalized Finsler geometry. The chapter also discusses such related issues as Fechner's original derivation of his logarithmic law, an observational version of the sorites paradox, a generalized Floyd–Warshall algorithm for computing metric distances from dissimilarities, an ultra-metric version of Fechnerian scaling, and data-analytic applications of Fechnerian scaling.

Gregory Ashby, Matthew J. Crossley, and Jeffrey Inglis review mathematical models of human learning in Chapter 4. Although learning was a key focus during the early years of mathematical psychology, the cognitive revolution of the 1960s caused the field to languish for several decades. Two breakthroughs in neuroscience resurrected the field. The first was the discovery of long-term potentiation and long-term depression, which served as promising models of learning at the cellular level. The second was the discovery that humans have multiple learning and memory systems that each require a qualitatively different kind of model. Currently, the field is well represented at all of Marr's three levels of analysis. Descriptive and process models of human learning are dominated by two different but converging approaches – one rooted in Bayesian statistics and one based on popular machine-learning algorithms. Implementational models are in the form of neural networks that mimic known neuroanatomy and account for learning via biologically plausible models of synaptic plasticity. Models of all these types are reviewed, and advantages and disadvantages of the different approaches are considered.

Marc W. Howard's Chapter 5 surveys formal models of memory. The idea that memory behavior relies on a gradually changing internal state has a long history in mathematical psychology. The chapter traces this line of thought from statistical learning theory in the 1950s, through distributed memory models in the latter part of the twentieth century and early part of the twenty-first century, through to modern models based on a scale-invariant temporal history. The author discusses the neural phenomena consistent with this form of representation and sketches the kinds of cognitive models that can be constructed with its use, in connection with formal models of various memory tasks.

In Chapter 6, Gregory Ashby and Michael Wenger review statistical decision theory, which provides a general account of perceptual decision-making in a wide variety of tasks that range from simple target detection to complete identification. The fundamental assumptions are that all sensory representations are inherently noisy and that every behavior, no matter how trivial, requires a decision. Statistical decision theory is referred to as signal detection theory (SDT) when the stimuli vary on only one sensory dimension, and as general recognition theory (GRT) when the stimuli vary on two or more sensory dimensions. SDT and GRT are both reviewed. The SDT review focuses on applications to the two-stimulus identification task and multiple-look experiments, and on response-time extensions of the model (e.g., the drift-diffusion model). The GRT review focuses on

applications to identification and categorization experiments, and in the former case, especially on experiments in which the stimuli are constructed by factorially combining several levels of two stimulus dimensions. The basic GRT properties of perceptual separability, decisional separability, perceptual independence, and holism are described. In the case of identification experiments, the summary statistics methods for testing perceptual interactions are described, and so is the model-fitting approach. Response time and neuroscience extensions of GRT are reviewed.

Chapter 7, written by Hans Colonius and Adele Diederich, deals with response inhibition, which is an organism's ability to suppress unwanted impulses, or actions and responses that are no longer required or have become inappropriate. In a stop-signal task experiment, participants perform a response time task (go task), and occasionally the go stimulus is followed by a stop signal after a variable delay, indicating subjects to withhold their response (stop task). The main interest of modeling is in estimating the unobservable latency of the stopping process as a characterization of the response inhibition mechanism. The authors analyze and compare the underlying assumptions of different models, including parametric and nonparametric versions of the race model. New model classes based on the concept of copulas are introduced, and a number of unsolved problems facing all existing models are pointed out.

In Chapter 8, written by Noah Thomas, Brandon M. Turner, and Trisha Van Zandt, approximate Bayesian analysis is presented as the solution for complex computational models where no explicit maximum likelihood estimation is possible. The activation-suppression race model (ASR), which does have a likelihood amenable to Markov chain Monte Carlo methods, is used to demonstrate the accuracy with which parameters can be estimated with the approximate Bayesian methods.

The cognitive diagnosis models considered in Chapter 9 by Jimmy de la Torre and Miguel A. Sorrel have their historical origins in the field of educational measurement, as a psychometric tool to provide finer-grained information suitable for formative assessment. Typically, but not necessarily, these models classify examinees as masters and nonmasters on a set of binary attributes. The chapter aims to provide a general overview of the original models and the extensions, and methodological developments, that have been made in the last decade. The topics covered in the chapter include model estimation, Q-matrix specification, model fit evaluation, and procedures for gathering validity and reliability evidences. The chapter ends with a discussion of future trends in the field.

Finally, Chapter 10, written by Fabian Soto and Gregory Ashby, reviews encoding models in neuroimaging. This is the neuroimaging area closest to mathematical psychology in which models of neuroimaging data are constructed by combining assumptions about underlying neural processes with knowledge of the task and the type of neuroimaging technique being used to produce equations that predict values of the dependent variable that is measured at each recording site (e.g., the fMRI BOLD response). Voxel-based encoding models include an encoding model

that predicts how every hypothesized neural population responds to each stimulus, and a measurement model that first transforms neural population responses into aggregate neural activity and then into values of the dependent variable being measured. Encoding models can be inverted to produce decoding schemes that use the observed data to make predictions about what stimulus was presented on each trial, thereby allowing unique tests of a mathematical model. Representational similarity analysis is a multivariate method that provides unique tests of a model by comparing its predicted similarity structures to similarity structures extracted from neuroimaging data. Model-based fMRI is a set of methods that were developed to test the validity of purely behavioral computational models against fMRI data. Collectively, encoding methods provide useful and powerful new tests of models – even purely cognitive models – that would have been considered fantasy just a few decades ago.

References

Aczel, J. A. (1987). *Short course on functional equations: Based upon recent applications to the social and behavioral sciences*. Dordrecht: Kluwer Academic.

Dzhafarov, E. (1995). Empirical meaningfulness, measurement-dependent constants, and dimensional analysis. In R. D. Luce, M. D'Zmura, D. Hoffman, G. J. Iverson, & A. Romney (Eds.), *Geometric representations of perceptual phenomena* (pp. 113–134). Mahwah, NJ: Lawrence Erlbaum Associates.

Falmagne, J.-C., Narens, L., & Doble, C. (2018). The axiom of meaningfulness in science and geometry. In H. Colonius, W. H. Batchelder & E. N. Dzhafarov (Eds.), *New Handbook of Mathematical Psychology: Modeling and measurement* (Vol. 2, pp. 374–456). Cambridge: Cambridge University Press.

Luce, R. D. (1959a). *Individual choice behavior*. New York: John Wiley & Sons.

Luce, R. D. (1959b). On the possible psychophysical laws. *Psychological Review*, *66*, 81–95.

Luce, R. D. (1962). Comments on Rozeboom's criticism of "on the possible psychophysical laws." *Psychological Review*, *69*, 548–551.

Narens, L. (2007). *Introduction to the theories of measurement and meaningfulness and the use of symmetry in science*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rozeboom, W. (1962). The untenability of Luce's principle. *Psychological Review*, 542–547.

1 Principles and Consequences of the Initial Visual Encoding

Brian Wandell and David Brainard

> Only infrequently is it possible to subject the manifold phenomena of life to simple and strict forms of mathematical treatment without forcing the data and encountering contradiction, probably never without a certain abandonment

> of the immense multiplicity of details to which those phenomena owe their aesthetic attractiveness. Nevertheless, however, it has often proved to be possible and useful to establish, for wide fields of biological processes and organic arrangements, comparatively simple mathematical formulas which, though they are probably not applicable with absolute accuracy, nevertheless simulate to a certain approximation a large number of phenomena. Such representations not only offer preliminary orientation in a field that at first seems completely incomprehensible, but they also often direct research into a correct course, in as much as first an insight into those fundamental formulations is sought, and then the deviations from their strict validity, which become apparent here and there, are made the subject of special investigations. Among the fields of physiology which have permitted the establishment of such guiding formulas the theory of visual sensations and of color mixture assumes a particularly distinguished position. (von Kries, 1902)

1.1 Introduction

Vision research has many purposes. Medical investigators aim to diagnose and repair visual disorders ranging from optical focus to retinal dysfunction to cortical lesions. Psychologists aim to identify and quantify the systematic rules of perception, including models of visual sensitivity, image quality, and the laws that predict percepts such as brightness, color, motion, size, and depth. Systems neuroscientists seek to relate visual experience and performance to the neural signals in the visual pathways, and computational investigators seek principles and models of perceptual and neural processes. Image systems engineers ask how to design sensors and processing to provide effective artificial vision systems.

Vision science draws upon findings from many fields, including biology, computer science, electrical engineering, neuroscience, psychology, and physics. Clear communication among people trained in different disciplines is not always straightforward. One of the ways that vision science has flourished is by using the language of mathematics to communicate core ideas. Vision science uses many types of mathematics; here we describe methods that have been used for many decades. These are certain linear methods, descriptions of noise distributions, and Bayesian inference. Many other linear methods (e.g., principal components, Fourier and Gabor bases, and independent components analysis) and nonlinear methods (e.g., linear–nonlinear cascades, normalization, information theory, and neural networks) can be found throughout the vision science literature. For this chapter, we focus on a few core mathematical methods and the complementary role of computation.

Physics – the field that quantifies the input to the visual system – provides mathematical representations of the light signal and definitions of physical units. The field of physiological optics quantifies the optical and biological properties of the lens. These properties are summarized as a mathematical transformation that maps the physical stimulus to the image focused on the retina, generally

referred to as the retinal image. At each retinal location the image is characterized as the spectral irradiance (power per unit area as a function of wavelength). Retinal anatomy and electrophysiology identify the properties of the rod and cone photoreceptors, enabling us to calculate the photopigment excitations from the retinal image using linear algebraic methods.

Perhaps the most famous use of mathematics in vision science is at the intersection of physics and psychology: the laws of color matching formalize the relationship between the physics of light and certain aspects of color appearance. The mathematical principles of color matching are also deeply connected to Thomas Young's biological insight that there are only three types of cone photopigment (Young, 1802). This insight implies a low-dimensional biological encoding of the high-dimensional spectral light. The linear algebraic techniques used to describe the laws of color matching were developed by the mathematician Hermann Grassmann. Indeed, he developed vector spaces in part for this purpose (Grassmann, 1853). The mathematics he introduced remains central to color imaging technologies and throughout science and engineering.

While acknowledging the importance of mathematical foundations, it is also important to recognize that there is much to be gained by building computational methods that account for specific system properties. The added value of computations is clear in many different fields, not just vision science. The laws of gravity are simple, but predicting the tides at a particular location on earth is not done via analytic application of Newton's formulas. Similarly, that color vision is three-dimensional is a profound principle, yet precise stimulus control requires accounting for many factors, such as variations of the inert pigments across the retinal surface (CIE, 2007; Whitehead, Mares, & Danis, 2006) and the wavelength-dependent blur of chromatic aberration (Marimont & Wandell, 1994). The mathematical principles guide, but we need detailed computations to predict precisely how color matches vary from central to peripheral vision.

We hope this chapter helps the reader value principles expressed by equations and computations embodied in software. Establishing the principles first provides a foundation for implementing accurate computations. Historically, our knowledge about vision has been built up by developing principles, testing them against experiments, and combining them with computation; this remains a useful and important approach. Indeed, we believe the goal of vision science includes not only producing models that account for performance and enable engineering advances, but also leveraging those models to extract new principles that help us think about how visual circuits work.

There are competing views: some would argue that large data sets combined with analyses using machine learning provide the best way forward to understanding, and recent years have seen impressive engineering advances achieved with this approach (D. D. Cox & Dean, 2014). We are certainly interested in the performance of such models as a point of departure, but here we emphasize principles and data-guided computational implementations of these principles.

This chapter begins by describing the representation of the visual stimulus, and how light rays in the scene pass through the optics of the eye and arrive at the retina. Next, we explain how the retinal photoreceptors (a) transform the retinal spectral irradiance into photoreceptor excitations, and (b) spatially sample the retinal image. Each of these steps can be expressed by a crisp mathematical formulation. To describe the real system with quantitative precision, we implemented software that models specific features of the scene, optics, and retina (ISETBio; Cottaris *et al.*, 2019, 2020; https://github.com/isetbio/isetbio/wiki), and we illustrate the use of these models in several examples.

The frontiers of vision science use mathematics to understand visual percepts, which provide a useful basis for thought and action. The information provided by light-driven photopigment excitations is used to create these percepts, but knowledge of the excitations alone falls far short of describing visual perception. The brain makes inferences about the external world from the retinal encoding of light, and throughout the history of vision science many investigators have suggested that the role of neural computation is to implement the principles that underlie these inferences. This point was emphasized as early as Helmholtz, who wrote:

> The general rule determining the ideas of vision that are formed whenever an impression is made on the eye, is that such objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism. (Helmholtz, 1866; English translation Helmholtz, 1896)

Within psychology this idea is called unconscious inference, a phrase that emphasizes that we are not aware of the neural processes that produce our conscious experience, an idea that was important to Helmholtz. Perhaps more important in this context is the principle that the percepts represent critical properties of external objects in the field of view, such as depth, reflectance, shape, and motion.

The mathematics of perceptual inference can take many forms, and in common scientific practice the mathematics of inference depend on what is known about the input signal. If the scene properties are not uniquely determined by the sensory measurements, such as when only three spectral classes of cones sample the spectral irradiance of the retinal image, probabilistic reasoning about the likely state of the world is inevitable. In vision science, linear methods combined with the mathematical tools of probabilistic inference are commonly used to understand how the brain interprets the mosaic of photoreceptor excitations to see objects, depth, and color. In the final part of this chapter we close the loop between sensory measurements and perceptual inference by introducing the mathematics of such inferences, focusing on two specific examples relevant to the study of the initial visual encoding. The principles we introduce, however, apply generally.

1.2 Scene to Retinal Image

1.2.1 Light Field

Light is the most important visual stimulus.[1] The word light means the electromagnetic radiation that is visible to the human eye.[2] The mathematical representation of light has been developed over many centuries through a series of famous experiments, and these experiments provide several different ways to think about light. Many properties of how light is encoded by the eye can be understood by treating light as comprising rays of many different wavelengths.

In a passage in his 1509 notebook (Da Vinci, 1970), Leonardo da Vinci noted that an illuminated scene is filled with rays that travel in all directions.[3] As evidence, he described a pinhole camera (camera obscura) made by placing a small hole in a wall of a windowless room (Figure 1.1). The wall is adjacent to a brightly illuminated piazza; an image of the piazza (inverted) appears on a wall within the room. Leonardo noted that an image is formed wherever the pinhole is placed, and he concluded that the rays needed to form an image must be present at all of these

Figure 1.1 *Light field geometry. The complete set of rays in the environment is the light field. The rays that arrive at the imaging system, in this figure a large pinhole camera, are the incident light field. If the imaging system includes a lens, rather than just a pinhole, the incident light field is described by the positions and angles of the rays at the lens aperture. Figure reproduced from Ayscough (1755).*

1 Mechanical force on the retina (pressure phosphenes) and injecting current into the retina or brain (electrical phosphenes) can also cause a visual sensation.

2 www.merriam-webster.com/dictionary/light

3 From the section prove how all objects, placed in one position, are all everywhere.

positions. Leonardo compared the space-filling light rays to the traveling waves that arise after dropping a rock in a pond.

The Russian physicist, Andrey Gershun, provided a mathematical representation of the geometry of these rays, which he called the light field (Gershun, 1939). The mathematical representation of the light field quantifies the properties of the light rays at each position in space [Equation (1.1)]. Each ray travels from a location (x, y, z) in a direction (α, β) and has a wavelength and polarization (λ, ρ). To know these parameters and the intensity of every ray is to know the light field at a given moment in time:

$$LF(x, y, z, \alpha, \beta, \lambda, \rho). \tag{1.1}$$

The light field representation does not capture some phenomena of electromagnetic radiation such as interference (waves) or the Poisson character of light (photon) absorption by the photoreceptors. Even so, the light field representation provides an excellent model to describe the ways in which light interacts with surfaces, and the geometric description of the light field is important in the mathematics of computer graphics, a technology that is important for illumination engineering, photography, and cinema (Pharr, Jakob, & Humphreys, 2016; Wald *et al.*, 2003, 2006).

1.2.2 The Incident Light Field

An eye – or a camera – records a small subset of the light field, those rays arriving at the pupil or entrance aperture. We call these the incident light field. In Figure 1.1 the dashed and solid lines are the light field and the solid lines are the incident light field. The natural parameterization of the incident light differs from the general light field. We can represent the incident light field using only the position (u, v) and angle (α, β) of the rays at the entrance aperture of the imaging system:

$$ILF(u, v, \alpha, \beta, \lambda, \rho, t). \tag{1.2}$$

Equation (1.2) also represents time (t) explicitly, which allows it to describe effects of motion both in the scene and by the eye.

1.2.3 Spectral Irradiance and the Plenoptic Function

The eye and most cameras do not measure the full incident light field. Rather, the rays are focused to an image at the retina or sensor, and the photodetectors respond to the sum across all directions of the image rays. To be explicit about this, Adelson and Bergen (1991) introduced the term plenoptic function, a simplified version of the incident light field, that was chosen to guide thinking about the computations carried out in the human visual pathways [their Equation (2)]. First, they approximated the eye as a pinhole camera; with this approximation all rays have the same entrance position **p**. Additionally, the retina/sensor surface defines the direction (**d**) of the rays that pass through the pinhole. For the pinhole case,

specifying two angles of a ray at the pinhole is equivalent to specifying the location where a ray will intersect the retina/sensor surface, (r_x, r_y). Finally, Adelson and Bergen ignored polarization as unimportant for human perception. With these restrictions, the plenoptic function for human vision is simply the retinal spectral irradiance, over time (t):

$$E(r_x, r_y, \lambda, t; \mathbf{p}, \mathbf{d}). \tag{1.3}$$

In Equation (1.3) we have explicitly reintroduced position and direction, but these are often implicit [as in the formulation of Equation (1.2) above]. Understanding the progression from light field to incident light field to retinal spectral irradiance is useful for understanding how the information available for visual processing relates to the complete set of potential information that could be sensed by a visual system.

Adelson and Bergen note that by placing the pinhole at many different positions and viewing directions, we can estimate the full light field from the set of spectral irradiances. It is possible to be more efficient and estimate the incident light field by using a lens, rather than a pinhole, inserting a microlens array over the photodetector array and placing multiple detectors behind each microlens. Both cameras and microscopes have employed this technology to support depth estimation (Adelson & Wang, 1992) and control focus and depth of field in post-processing (Ng *et al.*, 2005). Cameras that estimate the full incident light field are not currently in wide use (Wikipedia contributors, 2021); but, the widely used dual pixel autofocus technology obtains a coarse measure of the incident light field (Canon U.S.A., Inc., 2017; Mlinar, 2016). This is accomplished by inserting a microlens array over pairs of photodetectors. With this design rays from, say, the left and right sides of the lens are captured by adjacent detectors. This coarse estimate of the light field is useful for setting the lens focus and estimating depth.

1.2.4 The Initial Visual Encoding

Computational models of the early visual pathways define a series of transformations that characterize how the incident light field becomes a neural response. In this chapter, we introduce the mathematics used to characterize the initial visual encoding in the context of the first few of these transformations (Figure 1.2; see also Brainard & Stockman, 2010; Packer & Williams, 2003; Rodieck, 1998; Wandell, 1995). We focus on the encoding of the spectral radiance by the photoreceptors – subsequent neural processing operates on this visual encoding.

A visual scene's light field is generated by the properties and locations of the light sources and objects, and how the rays from the light sources are absorbed and reflected by the objects. Here we consider the special case of scenes presented on a flat display, so that in the idealized case where the display is the only object and there are no other light sources, the full light field is determined just by the spectral radiance emitted at each location of the display. Elsewhere, we consider

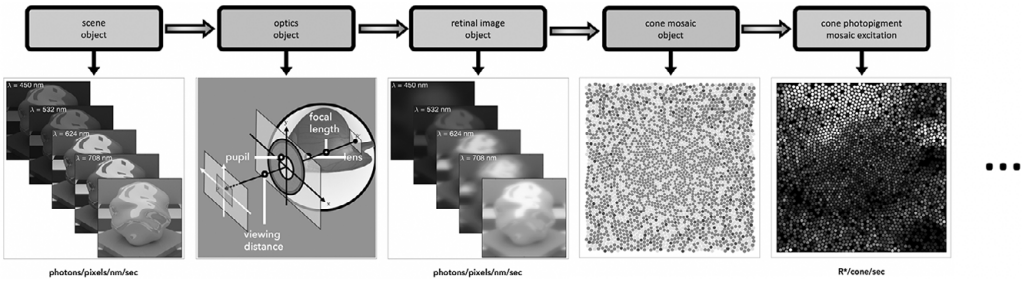


Figure 1.2 *The initial encoding of light by the visual system.* Scene*: An image on a display surface is characterized by the spectral radiance at each display location. Images of the display spectral radiance are shown at a few sample wavelengths, along with a rendering of the image.* Optics*: The incident light field enters the pupil of the eye and a spectral irradiance image is formed on the retina. The* retinal image *is blurred relative to the displayed image, and the spectral irradiance is affected by lens and macular pigment absorptions.* Cone mosaic*: The retinal image is spatially sampled by the L-, M-, and S-cone mosaics.* Cone excitations*: The retinal image irradiance, spectrally weighted by each cone photopigment absorptance function, is integrated within the cone's aperture and temporally integrated over the exposure duration to produce a pattern of cone excitations. This figure should be viewed in color. The color version is available at https://color.psych.upenn.edu/supplements/earlyencoding/computationsColorFig.pdf. We thank Nicolas Cottaris for the figure.*

the more general case of modeling the formation of the retinal spectral irradiance, given a description of the light sources and objects in a three-dimensional scene (Lian *et al.*, 2019).

The optics of the eye collect the incident light field and focus the rays to produce the spectral irradiance arriving at the retina. Factors such as diffraction and aberrations in the eyes optics mean that this image is blurred relative to the displayed image. In addition, wavelength-selective absorption of short-wavelength light by the lens and inert macular pigment also affect the spectral irradiance. Of note (but not illustrated in Figure 1.2), the density of the macular pigment is high in the central area of the retina and falls off rapidly with increasing eccentricity.

Photoreceptors spatially sample the retinal image. Excitations of photopigment molecules in these photoreceptors provide the information available to the visual system for making perceptual inferences about the scene. Here we consider the cone photoreceptors, which operate at light levels typical of daylight. There are three spectral classes of cones, each characterized by its own spectral sensitivity. That there are three classes leads to the trichromatic nature of human color vision. Figure 1.2 illustrates a patch of cone mosaic from the central region of the human retina. The properties of the mosaic are quite interesting. For example, there are no S-cones in the very center of the retina, and many properties of the

mosaic (e.g., cone density, cone size, cone photopigment optical density) vary systematically with eccentricity (Brainard, 2015; Hofer & Williams, 2014).

Not considered here is a separate mosaic of highly sensitive rod photoreceptors that is interleaved with the cone mosaic. The rods mediate human vision at low light levels (Rodieck, 1998). We also ignore the melanopsin containing intrinsically sensitive retinal ganglion cells (Gamlin *et al.*, 2007; Hattar *et al.*, 2002; Van Gelder & Buhr, 2016). The principles we develop, however, also apply to modeling the excitations of these receptors.

Modeling of the initial visual encoding is well understood, and we explain the key linear systems principles next, using a simplified representation of the light stimulus. Advanced modeling of the subsequent neural processes includes non-linearities; the mathematical principles and computational methods we introduce are a fundamental part of the full description. After explaining the mathematical principles, we illustrate how to extend them through computational modeling that harnesses the power of computers to characterize biological reality in more detail than is possible with analytic calculations alone.

1.3 Mathematical Principles

1.3.1 Linear Systems

Linear systems and the tools of linear algebra are the most important mathematical methods used in vision science. Indeed, when trying to characterize a system, the scientist's and engineer's first hope is that the system can be approximated as linear. A system, L, is linear if it follows the superposition rule:

$$L(x + y) = L(x) + L(y). \tag{1.4}$$

Here x and y are two possible inputs to the system and $x + y$ represents their superposition. The homogeneity rule of linear systems follows from the superposition rule. Consider that

$$\begin{aligned} L(x + x) &= L(2x) \\ &= L(x) + L(x) \\ &= 2L(x). \end{aligned}$$

This is easily generalized for any integer m to show that:[4]

$$L(mx) = mL(x). \tag{1.5}$$

4 It is an exercise for the reader to show that a system that follows the superposition rule also obeys the homogeneity rule, not just for integers, but for any real scalar. If x is a real-valued scalar, homogeneity also implies superposition. When $\mathbf{x}$ is a real-valued vector with entries x_n, however, a system can obey homogeneity but not superposition. For example, $f(\mathbf{x}) = \sqrt[3]{\sum x_n^3}$ satisfies homogeneity but not superposition. The reader may find it of interest to consider why we used an exponent of three rather than two for this example.

No physical system can be linear over an infinite range – if you put enough energy into a system it will blow up! But many systems are linear over a meaningful range of input values.

1.3.2 Linearity Example: Cone Excitations and Color Matching

Vision is initiated when a photopigment molecule absorbs a photon of light. The absorption can cause the photopigment, a protein, to change conformation, an event we refer to as a photopigment excitation. The excitation initiates a molecular cascade inside the photoreceptor that changes the ionic currents at the photoreceptor membrane. The change in current modulates the voltage at the photoreceptor synapse and causes a release of neurotransmitter (Rodieck, 1998).

The transformation from the spectral energy of light, $E(\lambda)$, incident upon a cone to the number of photopigment excitations, n, produced by that light is an important, early vision, linear system. Consider two different spectra, denoted by $E_1(\lambda)$ and $E_2(\lambda)$. Let L represent the system that describes the transformation between spectra and excitations. This system obeys the superposition rule:

$$L(E_1 + E_2) = L(E_1) + L(E_2). \tag{1.6}$$

This linearity holds well over a wide range of light levels typical of daylight natural environments (Burns *et al.*, 1987).

An important feature of photopigment excitations is that their effect on the membrane current and transmitter release does not differ with the wavelength of the exciting photon. Such differences might have existed because different wavelengths are preferentially absorbed at different locations within the cone outer segment, or because photons of different wavelengths carry different amounts of energy. The observation that all excitations have the same impact is called the Principle of Univariance. As Rushton wrote:

> The output of a receptor depends upon its quantum catch, but not upon what quanta are caught. (Rushton, 1972)

The color-typical human retina contains three distinct classes of cones, which are referred to as the L (long-wavelength sensitive), M (middle-wavelength sensitive), and S (short-wavelength sensitive) cones. While the effects of photopigment excitations are univariant, the probability of a photopigment excitation is wavelength-dependent. The wavelength-dependent probability that an incident photon leads to an excitation is characterized by the pigment's spectral absorptance.[5] The absorptance depends on the density of the photopigment within the

5 The absorptance spectrum is the probability that a photon is absorbed. Not all absorbed photons lead to an excitation, so an additional factor specifying the quantal efficiency (probability of excitation given absorption) needs to be included in the calculation. Current estimates put the quantal efficiency of human cone photopigment near 67%. In addition, the calculation of cone excitations from spectral irradiance requires taking into account the size of the cone's light-collecting aperture.

cone's outer segment, as well as on the outer segment length; details are elaborated elsewhere (Rodieck, 1998; see also Packer & Williams, 2003; Pugh, 1988).

It is difficult to measure the light at the retinal surface in the living eye, but it is straightforward to measure the light incident at the cornea. Hence, it is typical to specify the absorptance with respect to the spectrum of the light incident at the cornea. This convention effectively combines the effects of the lens, the inert retinal macular pigment, the photopigment absorptance, and quantal efficiency. For simplicity, vision scientists call the cornea-referred spectral excitation curve the cone fundamental.

The three (L-, M-, and S-) cone fundamentals define for each cone type the probability of excitation given the spectrum of light entering the eye. The human cone fundamentals have been carefully measured and tabulated (Figure 1.3; Stockman & Sharpe, 2000; Stockman, Sharpe, & Fach, 1999; www.cvrl.org) and are the subject of an international standard (CIE, 2007).

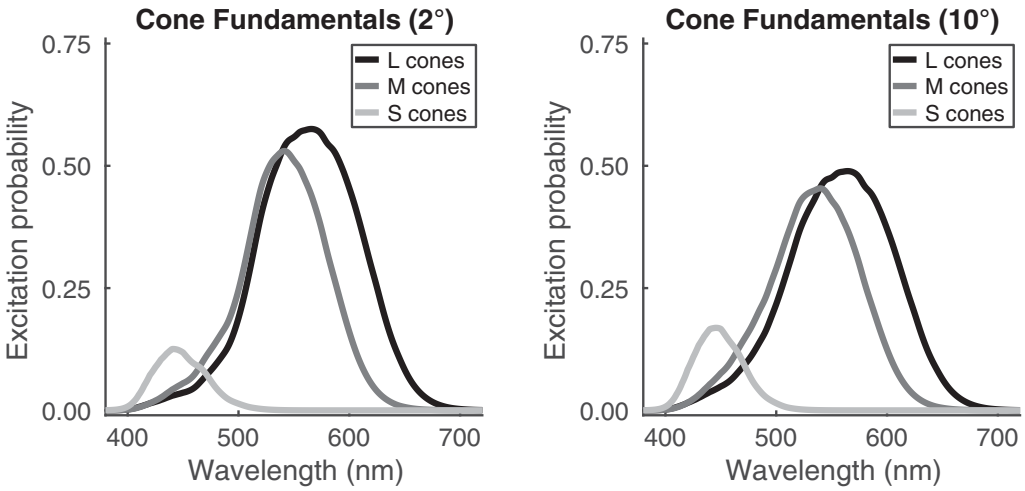


Figure 1.3 *Human cone fundamentals. The left panel shows estimates of the L-, M-, and S-cone fundamentals for foveal viewing. The fundamentals are the probability of excitation per photon entering the cone's entrance aperture, but with pre-retinal absorption taken into account. Note the large difference between the L- and M-cone fundamentals compared to the S-cone fundamental. This difference is due partly to the selective absorption of short-wavelength light by the lens and macular pigment. The right panel shows estimates for cones at 10° eccentricity. The S-cone fundamental is relatively higher at 10°, because there is little or no macular pigment at that eccentricity; and for the same reason there is a slight change in the relative values of the L- and M-cone fundamentals. In addition, the cone outer segment lengths decrease with eccentricity, leading to the lower peak probability of excitation in the periphery. This reduction, however, is more than compensated for by an increase in the size of the cone apertures with eccentricity. The impact of the aperture is not shown in these plots, but see Figure 1.4.*

To compute the number of cone excitations we use linear formulas. Suppose that a cone's fundamental is given by $C(\lambda)$. Using linearity and continuous mathematics, we compute the number of excitations at a single location as

$$N(r_x, r_y) = \int C(\lambda)E(r_x, r_y, \lambda)d\lambda. \tag{1.7}$$

The discrete form of this integral, commonly used in computational methods, is the inner product of the cone fundamental with the cornea-referred spectral irradiance incident upon a retinal location:[6]

$$N(r_x, r_y) = \sum_{\lambda_i} C(\lambda_i)E(r_x, r_y, \lambda_i)\Delta\lambda. \tag{1.8}$$

Here the λ_i are a set of w discretely sampled wavelengths, and $\Delta\lambda$ is the wavelength sample spacing.

1.3.3 Matrix Formulation of Linearity

We can calculate cone excitations by a matrix multiplication. The matrix $\mathbf{C}$ combines the three discretized cone fundamentals $C_L(\lambda_i)$, $C_M(\lambda_i)$, and $C_S(\lambda_i)$ into its rows, so that its dimension is $3 \times w$. Similarly, we write the spectral irradiance at a position, $E(r_x, r_y, \lambda)$, as a $w \times 1$ vector $\mathbf{e}(r_x, r_y)$. The L-, M-, and S-cone excitations available at a retinal location are described by a three-dimensional column vector:

$$\mathbf{n}(r_x, r_y) = \mathbf{C}\mathbf{e}(r_x, r_y). \tag{1.9}$$

The vector field $\mathbf{n}(r_x, r_y)$ describes the potential information available to the visual system from the cones at a moment in time. This representation replaces the dependence of the spectral irradiance on wavelength with the excitations of the three classes of cones. As we describe in more detail below, not all of this potential information is sensed by the visual system, since the cones discretely sample $\mathbf{n}(r_x, r_y)$.

It is worth reflecting on the implication of the linearity expressed by Equation (1.9). If we measure the cone fundamentals at each of the sample wavelengths λ_i, we can predict the cone excitations to any spectrum $E(r_x, r_y, \lambda_i)$. Thus, linearity implies that we can compute the system response to any input after making enough measurements to determine the system matrix $\mathbf{C}$. The ability to delineate the set of measurements required for complete system characterization is an important consequence of linearity, and this observation applies to linear systems in general, not just to computation of cone excitations.

A second implication of Equation (1.9) concerns which spectral radiances appear to be the same; these pairs are called metamers. Young (1802) had proposed that metamers arise if two lights produce the same set of cone excitations. This

6 In this formulation, we do not make the spatial extent of the cone acceptance aperture explicit. This aperture introduces additional blur into the retinal image. Computational models (see Figure 1.7) account for this factor; it is significant.

implies that the difference between a metameric pair is in the null space of the matrix, **C**.[7] That is, $\mathbf{e_1}$ and $\mathbf{e_2}$ must satisfy

$$\begin{aligned} \mathbf{Ce_1} &= \mathbf{Ce_2} \\ \mathbf{0} &= \mathbf{C}(\mathbf{e_2} - \mathbf{e_1}). \end{aligned} \tag{1.10}$$

Wyszecki (1958; see also Wyszecki & Stiles, 1982) referred to vectors in the null space of **C** as metameric black spectra. Adding a nonzero metameric black to any spectrum produces a metamer.

Displays and printers do not reproduce the original physical stimulus; rather, they create lights designed to be metamers to the original. Thus, calculating metamers is central to color reproduction technologies. Practical aspects of the computation of metamers for color reproduction applications, including limitations based on the spectra a device can produce, are discussed in detail elsewhere (Brainard & Stockman, 2010; Hunt, 2004).

1.3.4 Color-Matching Functions

James Clerk Maxwell (1860) was the first to measure pairs of spectral irradiance functions, $\mathbf{e_1}$ and $\mathbf{e_2}$, that appear the same to humans despite being physically different. These data place constraints on estimates of the matrix **C**, but do not uniquely determine it. To understand why, note that the null space of **C** is the same as the null space of $\mathbf{T} = \mathbf{MC}$, for any invertible 3×3 matrix **M**. Thus, any such matrix **T** predicts the same set of matches.

The rows of **T**, when viewed as functions of wavelength, are referred to as a set of color-matching functions. We say that the color-matching functions are only unique up to a linear transformation. The technology for creating metamers relies on color-matching functions which were chosen as an international standard (CIE, 1986, 2007). How color-matching functions may be obtained directly from perceptual color-matching experiments, without explicit reference to the cone fundamentals, is treated in many sources (Brainard & Stockman, 2010; Wandell, 1995; Wyszecki & Stiles, 1982). Indeed, high-quality measurements of behavioral color matching (e.g., Stiles & Burch, 1959) provide key data that constrain modern estimates of human cone fundamentals.

There are a number of properties of the eye that must be modeled if we are to compute a true estimate of cone mosaic excitations. For example, only one type of cone is present at each position, so we must specify a cone spatial sampling scheme. That is the reason that we use the term *potential information* to describe cone excitations $\mathbf{n}(r_x, r_y)$ as a function of retinal location – not all of that information is sampled by the cone mosaic. Also, as noted above, the density of both inert pigments and photopigments varies with retinal location, as does the size of the

7 The null space of a matrix **C** is the space of vectors **v** such that $\mathbf{Cv} = 0$. If a matrix has column dimension n and rank r, its null space has dimension $n - r$.

cone apertures. Enough is known about these properties to enable us to compute a reasonable approximation to the cone mosaic excitations across the retina.

1.3.5 Noise in the Sensory Measurements

Measurement noise is fundamental in the physical sciences and engineering. Two types of noise are used throughout the sensory sciences: Gaussian (normal) noise and Poisson noise. Gaussian noise has two parameters (a mean and variance) but the Poisson distribution has a single parameter (the Poisson mean equals its variance). The formulas for the Gaussian density function and Poisson probability mass function, along with example draws from these distributions, are provided in Figure 1.4.

The Gaussian and Poisson distributions can be compared by setting the Gaussian mean equal to its variance. For small values, the Gaussian has values below zero. As the Poisson mean increases, the matched Gaussian is extremely similar (Figure 1.4).

There is an important conceptual difference between how these noise distributions are used in applications. There are many theorems about additive Gaussian noise, and thus it is common to introduce noise in a model with such noise using a fixed mean (μ) and standard deviation (σ). The added noise has the same distribution for all values of the signal (signal-independent noise).

For typical sensor measurements, including the cone excitations, the noise depends on the signal. Specifically, for the cones and many other measurement devices, the noise is Poisson distributed, with the Poisson parameter equal to the mean number of excitations (signal-dependent noise). The difference between signal-independent and signal-dependent noise can be quite significant (Figure 1.4).

1.3.6 Image Formation

The linear system principles described for one-dimensional spectral functions can be extended to two-dimensional functions, such as images. We use linear system methods to analyze how the cornea and lens form the retinal image. An important, but simple, case occurs for an image confined to a plane, such as a visual display or an optometrist's eye chart. For such images we can estimate the spectral irradiance at the cone apertures using a two-dimensional linear system computation.

The image emitted from a visual display is a function of position (x, y) and wavelength λ. As a first approximation, the display emits the same density of rays over a wide angle, which is why the display appears to be approximately the same when seen from different positions. The image from the display is called the spectral radiance, $I(x, y, \lambda)$, and it has units of W/sr/m^2/nm.

The spectral irradiance at the retina, $E(r_x, r_y, \lambda)$, is formed from the cone of rays that are captured by the pupil. In this case, r_x and r_y specify retinal location and the units of the image are those of spectral irradiance, W/m^2/nm, which result from integration over the solid angle of the pupil.

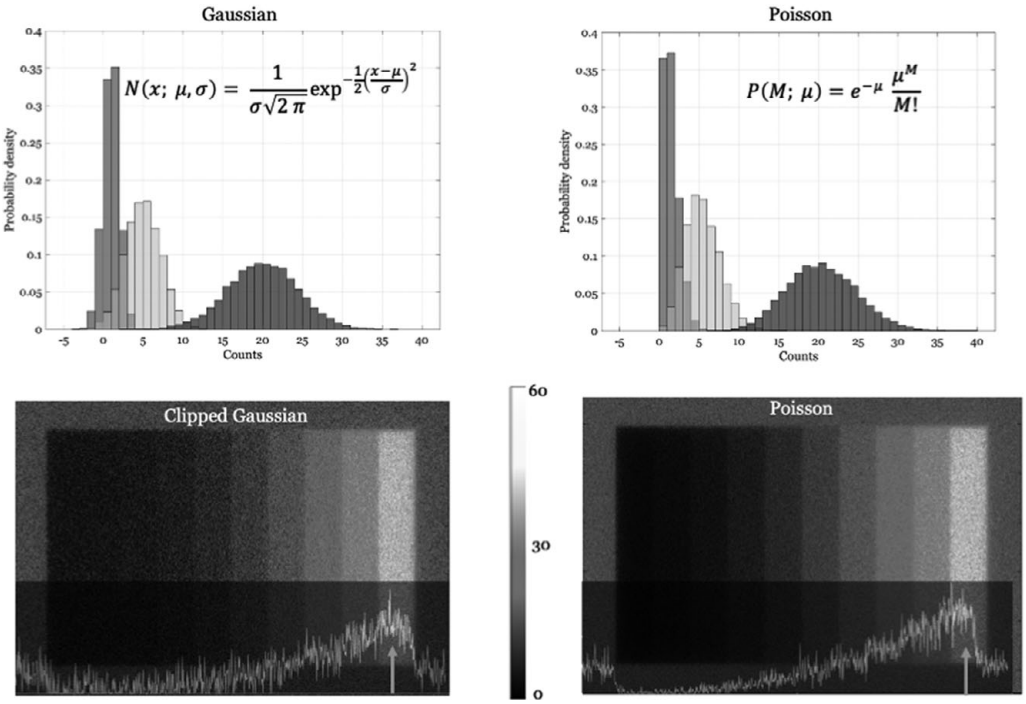


Figure 1.4 *The number of cone excitations is inescapably noisy, following a signal-dependent Poisson distribution. (Top) For mean values greater than 10, the Poisson distribution is reasonably approximated by a Gaussian distribution with a mean equal to the variance. For smaller values, it is necessary to clip the negative values for the Gaussian to achieve a good approximation. Low excitation rates are common under low-light conditions and for nearly all conditions when assessing the S-cones and rods (Baylor* et al.*, 1979; Hecht, Schlaer, & Pirenne, 1942). (Bottom) The signal-dependent nature of Poisson noise is important; simply adding Gaussian noise with a fixed mean is not a good approximation if there is a substantial range in the mean excitation values. The images illustrate the excitations in response to a series of bars spanning a large range of mean excitation using a signal-independent clipped Gaussian noise (left) and a Poisson noise (right). The Gaussian distribution added to the signal has zero mean and variance equal to the number of excitations in the brightest bar (arrows); this approximates Poisson noise for that bar. The inset trace, which shows excitations across a row of the image, illustrates that the Gaussian noise is too large for the dark bars. Had the variance been set to match the noise at the dark bar, the clipped Gaussian would be too small for the brightest bar. The simulation was created for an array of M-cones in the central fovea, a 2 ms exposure duration, achromatic bars of increasing intensity, and a bright bar luminance of 300 cd/m^2. This figure should be viewed in color. The color version is available at https://color.psych.upenn.edu/supplements/earlyencoding/noiseColorFig.pdf.*

The key linear system idea [Equation (1.6)] holds for retinal image formation (Wandell, 1995). If two input images, $I_1(x, y, \lambda)$ and $I_2(x, y, \lambda)$, produce two retinal images, $E_1(r_x, r_y, \lambda)$ and $E_2(r_x, r_y, \lambda)$, then the superposition of the input images, $I_1(x, y, \lambda) = I_1(x, y, \lambda) + I_2(x, y, \lambda)$, produces the superposition of the retinal images:

$$E(r_x, r_y, \lambda) = E_1(r_x, r_y, \lambda) + E_2(r_x, r_y, \lambda). \tag{1.11}$$

It follows that if the input image is the weighted sum of two input images, $I(x, y, \lambda) = \alpha I_1(x, y, \lambda) + \beta I_2(x, y, \lambda)$, the output retinal image will be the weighted sum of the two corresponding retinal images:

$$E(r_x, r_y, \lambda) = \alpha E_1(r_x, r_y, \lambda) + \beta E_2(r_x, r_y, \lambda). \tag{1.12}$$

As noted above, an important consequence of linearity is that it tells us how to generalize. When we know the response to an image I_k, measuring the response to a second image, I_j, enables us to predict the responses to an entire class of new images, all images of the form $\alpha I_k + \beta I_j$.

1.3.7 Shift-Invariance and Convolution

To characterize color matching we used the fact that a discrete linear system may be expressed as a matrix multiplication [Equation (1.9)]. A matrix can also be used to express retinal image formation, but in this case the number of measurements required to determine the requisite matrix is very large. For this reason, we consider an additional special and simplifying property linear systems can have: shift-invariance. These are linear systems such that shifting the position of the input correspondingly shifts the position of the output, without changing its form.[8] It is possible to measure whether a system is shift-invariant by a simple experiment. For an input image, say $I(x, y, \lambda)$, measure the retinal image $E(r_x, r_y, \lambda)$. Then shift the input, $I(x - \delta x, y - \delta y, \lambda)$, and measure the retinal image again. If for all choices of $(\delta x, \delta y)$ in the image domain, the output is shifted equivalently, $E(r_x - \delta r_x, r_y - \delta r_y, \lambda)$ in the retinal image domain, then the system is shift-invariant. Here the retinal image shifts $(\delta r_x, \delta r_y)$ differ from their image counterparts $(\delta x, \delta y)$ by the factor that converts the positional units of the image to those of the retinal image.

We can express linearity and shift-invariance using the convolution formula. For simplicity, we choose one wavelength and suppress λ. Suppose $P(r_x, r_y)$ is the retinal image from an image that is just a single point. The image $P(r_x, r_y)$ plays a central role in the characterization of convolutional optical systems: it is called the point spread function.[9] The point spread function is all we need to compute the retinal image for any input image. The idea is to treat the input image as a set of points, and to add shifted copies of the point spread function, each weighted by the input image intensity:

8 When describing optics, a shift-invariant region within the visual field is called an isoplanatic region.

9 The point spread function is the spatial analog of the impulse response function used to characterize time-invariant linear systems.

$$E(r_x, r_y) = \int_u \int_v I(u,v)P(r_x - u, r_y - v)dudv. \quad (1.13)$$

The importance of linear shift-invariance is that we characterize the system fully by one measurement, $P(r_x, r_y)$. We use the convolution formula and this measurement to compute the responses of a linear shift-invariant system to any input.

While shift-invariance and convolution are important concepts, the eye's optics deviate significantly from this ideal. Shift-invariance is a good approximation of human retinal image formation in local regions, say spanning a few degrees of visual angle and a change in wavelength of 20–50 nm. Properties of the photoreceptor sampling mosaic further limit the accuracy of the shift-invariant approximation of the visual encoding (see Figure 1.6). Thus the convolutional approximation is helpful for thinking about encoding over small regions, but it is not an accurate depiction when one considers a larger field of view. A realistic approximation requires computational modeling.

1.4 Computational Model of the Initial Encoding

The mathematical principles described above tell us how to compute the retinal image and the noisy cone excitations from a displayed image; the calculations are straightforward for a single retinal location. But an accurate model of the visual system must account for variations in the optics, pigments, and sampling properties of the cone mosaic with visual field location. These are substantial and impact the information available to the brain for making perceptual inferences about the visual scene. Parameters with significant spatial variation across the visual field include the optical point spread function, density and size of the cones in the mosaic, the distribution of different cone types within the overall mosaic, and the cone fundamentals. To make a realistic calculation requires implementing a computational model of the visual transformations.

1.4.1 The Value of Computational Modeling

Carefully validated computer simulation of the initial visual encoding has the potential to support advances in understanding many aspects of visual function. We use image-computable models to build upon the mathematical characterizations – earned through 400 years of experimental and theoretical work in vision science – and estimate the initial visual signals. Such knowledge is an essential foundation to use when modeling less well understood visual processes. The models help us separate effects attributable to known factors of the initial encoding from effects of factors that arise in later processing. For example, understanding cortical visual processing requires representing the input to the cortex. Without accurate modeling of the input, we risk attributing features of the cortical signals to the wrong neural mechanisms.

Because of the central role computational modeling plays in understanding vision, we have invested in developing a set of freely available software tools to model retinal image formation and cone excitations (Image Systems Engineering Tools for Biology – ISETBio; https://github.com/isetbio/isetbio.git; Cottaris *et al.*, 2019, 2020). The tools can be used for images presented on planar displays and for full three-dimensional descriptions of the objects and light sources in the scene (Image Systems Engineering Tools 3D; https://github.com/iset/iset3d.git; Lian *et al.*, 2019). In this section we briefly illustrate some basic calculations enabled by ISETBio. We are not advocating for our implementation in particular, but we do believe that the field needs to develop trusted open-science tools for computational modeling.

1.4.2 Shift-Varying and Wavelength-Dependent Point Spreads

The point spread functions from a single subject, measured at different retinal locations and wavelengths, differ significantly (Figure 1.5). The variation with retinal location occurs because the optical aberrations depend on the direction of the rays incident at the retina. The ISETBio tools can explicitly represent the full

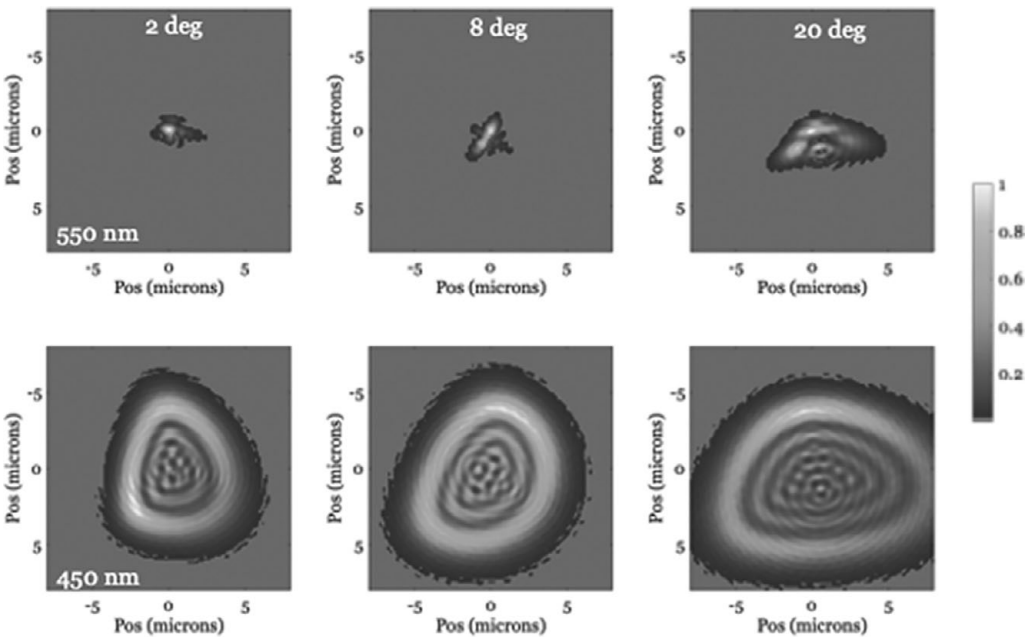


Figure 1.5 *The human point spread function. The images in the top row show the point spread functions at 550 nm from a typical subject measured at three different visual eccentricities. The point spread increases with eccentricity. The bottom images show the point spread but for light at 450 nm. The human eye cannot focus these two wavelengths at the same time because the index of refraction in the lens and cornea is wavelength-dependent. For many people, chromatic aberration is the largest aberration. A diagram showing simple ways to estimate degrees of visual angle is available from Branwyn (2016).*

incident light field and calculate these effects from a model eye (Lian *et al.*, 2019). Improvement of eye models is an active area of investigation, and in some cases ISETBio relies on empirical measurements of the eye's optics to predict responses over a range of retinal field locations (Jaeken & Artal, 2012; Polans *et al.*, 2015).

The point spread function varies with pupil diameter and wavelength in addition to visual field position. The dependence on pupil diameter, which varies with the light level of the scene, occurs for two reasons. As the pupil opens, the aberrations vary because more of the imperfectly shaped corneal and lens surfaces refract the light. As the pupil closes, diffraction starts to be a significant factor. The wavelength dependence is explained by the refractive indices of the cornea and lens. These chromatic aberrations are the largest of all the aberrations (Thibos *et al.*, 1990; Wandell, 1995).

1.4.3 Shift-Varying Sampling

Figure 1.6 shows the spatial arrangement of cones at different locations within the retina. The cone density is highest in the central fovea where the cones are tightly packed. Moving away from the center, cone density falls off and the cone apertures become larger. As cone density decreases, rod photoreceptors (the smaller receptors in the peripheral images) appear and fill the gaps between the cones. In addition, not apparent in the figure, cones become shorter away from the fovea. The shortening reduces the spectral absorptance.

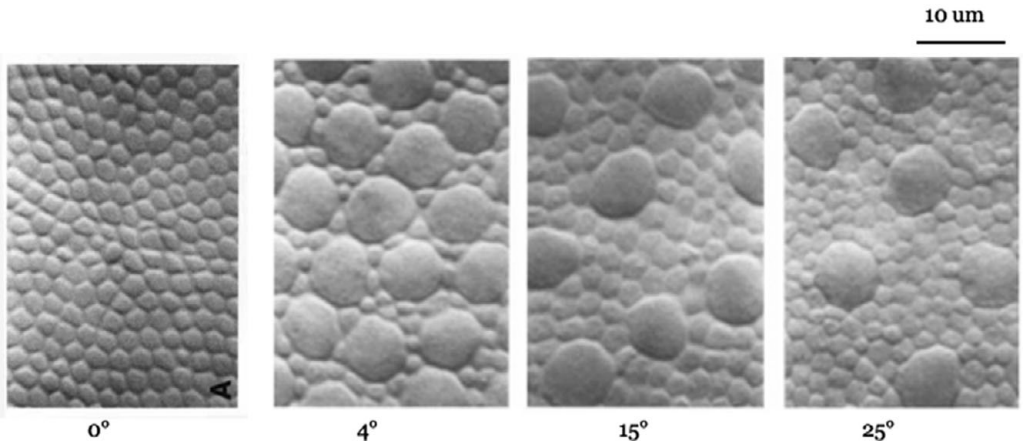


Figure 1.6 *Human cone and rod sampling mosaics. The en face images show the photoreceptor inner segments, where light enters the cones, at four retinal eccentricities. In the central region, all of the receptors are cones. At 4° and beyond, the large apertures are the cones and the smaller apertures are the rods. The cone sampling density and cone aperture sizes differ substantially between the central fovea and other visual eccentricities. The reduced sampling density limits the spatial resolving power of the eye. The larger cone apertures increase the rate of photon excitations per cone. Scale bar is 10 μm. Recomposited from figures in Curcio* et al. *(1990).*

The sampling density reduction means that less spatial information about the retinal image is extracted at retinal locations away from the central fovea. The relative density of the different cone types also varies with eccentricity. Indeed, as noted above, there are no S-cones in the very central fovea (Williams, MacLeod, & Hayhoe, 1981), so that vision in this small retinal region is dichromatic rather than trichromatic. Perhaps this region is specialized for high-resolution vision and omitting a few S-cones, which see a blurry retinal image at short wavelengths because of the chromatic aberrations, maximizes the information transmitted to the brain about spatial structure (Brainard, 2015; Garrigan *et al.*, 2010; Hofer & Williams, 2014; Williams *et al.*, 1991; Zhang, Cottaris, & Brainard, 2021).

The impact of the cone size and density, along with variations in the inert pigments described above, mean that calculating the cone excitations is shift-varying: the calculation is linear, but the parameters change with eccentricity. These eccentricity-dependent calculations are included in the ISETBio simulations. There is little value in expressing the full complexity of these calculations in pure mathematical form.

The impact of the several eccentricity-dependent factors on the cone excitations is substantial and illustrated in Figure 1.7. The images in the left column illustrate calculations in the central fovea and the images in the right column illustrate the same calculations at 10° in the periphery. The top image shows the differences in the size and density of the cone photoreceptor apertures. Also, notice the absence of S-cones in the small region of the very central fovea. The images inset in the top show the size of the point spread function for an in-focus wavelength: there are many more cones within the foveal point spread than within the 10° point spread.

The images in the middle row represent the number of cone excitations in response to a relatively low-frequency grating pattern. There are more excitations per cone at 10° than in the fovea, and there are many more cones representing the stimulus in the fovea. The third row shows the effect of increasing the stimulus spatial frequency. The foveal mosaic samples densely enough to preserve the regular pattern, but at 10° the spatial samples look like a wobbly representation of the stimulus.

Finally, notice that many cones have relatively low excitation levels to this achromatic stimulus. These cones appear as the quasi-regular array of black dots that are easy to see at 10°. They are also present, but harder to see, in the excitations for the central location. These cones are the S-cones, which absorb many fewer photons than the L- and M-cones. This lower excitation rate is partly due to the spectral transmission of the lens (and in the central region the macular pigment), which absorbs a great deal of short-wavelength light.

In summary, the principles of linearity and shift-invariance are useful guides for reasoning about cone excitations. These principles were part of our toolkit as we built a specific model of the human eye, and so they would be for any model. However, in the human eye deviations from shift-invariance are substantial. In addition, there are significant differences between people that may be important for explaining between-subject differences. Thus, an essential ingredient for building

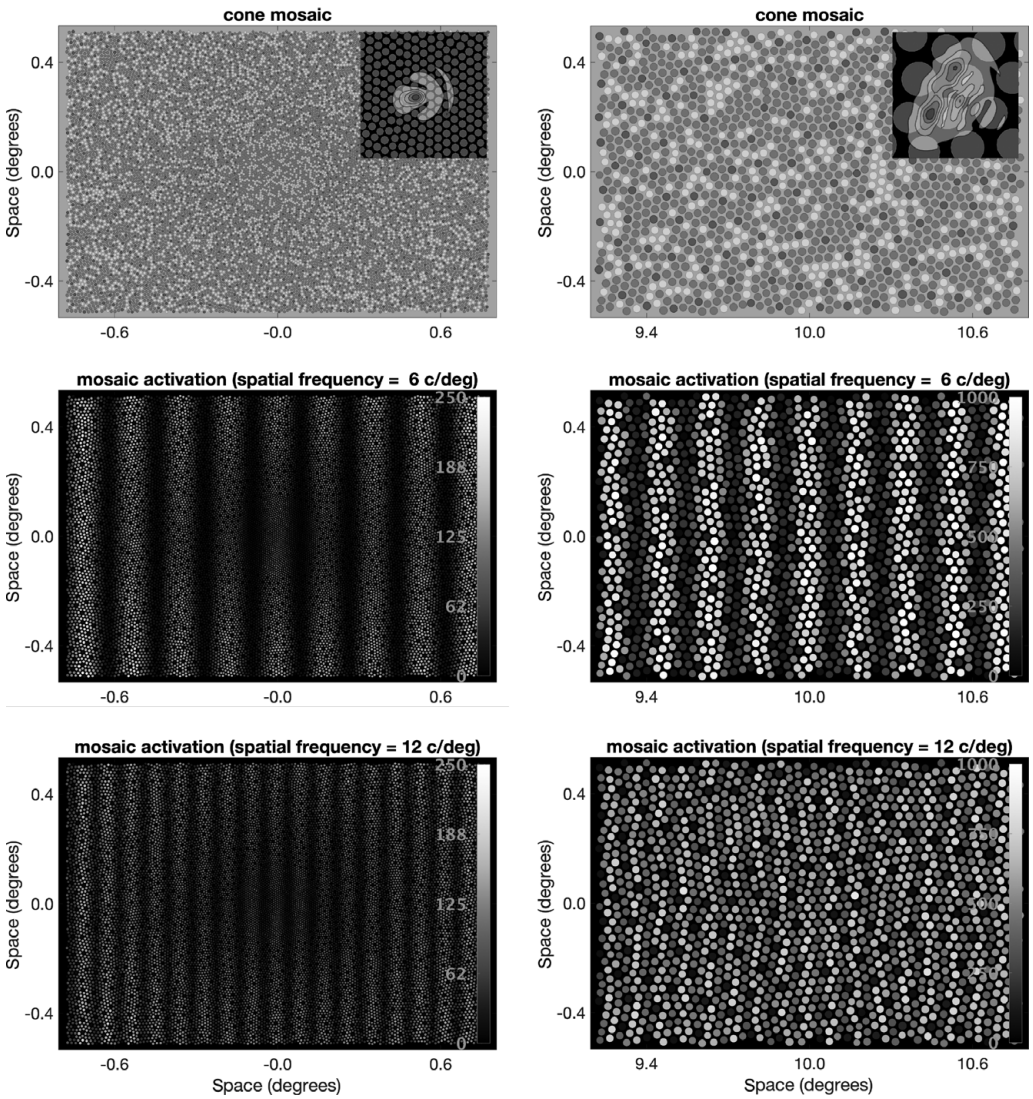


Figure 1.7 *Excitation calculations. The two columns represent two retinal eccentricities, each about* $1^{\circ 2}$. *(Top) The interleaved L-, M-, and S-cone mosaics, shown as red, green, and blue dots, are shown at the top. The inset shows an expanded view of the point spread function in the same region. The rods are not represented. (Middle and bottom) The gray level in these images shows the estimated cone excitations for a 6 c/deg harmonic and a 12 c/deg harmonic. The scale for the foveal location runs between 0 and 250, while that for the peripheral location runs from 0 to 1000. Peripheral cones have more excitations to the same stimulus because the cone apertures are larger. This figure should be viewed in color. The color version is available at https://color.psych.upenn.edu/supplements/earlyencoding/excitationsColorFig.pdf. Figure courtesy of Nicolas Cottaris.*

a computational model are data sets that quantify the critical model parameters (e.g., how the optical point spread function and cone density vary with visual field position) and how these parameters vary across individuals. For these reasons, a computational model is essential for applications that aim to create realistic estimates of the cone excitations for a population.

The computational implementation has benefited from mathematical principles and from data collected and shared by many investigators. Conversely, the exercise of building computational models often highlights the need for data sets that do not yet exist (e.g., across individuals, are optical quality and cone density independent, or do they covary in some systematic way?) At this point in the chapter, the reader might find it useful to re-read the quote at the start of this chapter, which was written by von Kries, Helmholtz's greatest disciple (Cahan, 1993), more than a century ago.

1.4.4 Spatial Derivatives of the Cone Excitations Mosaic

Adelson and Bergen (1991) observed that the partial derivatives of the spectral irradiance correspond to computations performed by neurons in the early visual system. Figure 1.8 illustrates these derivatives for several cases: derivatives with

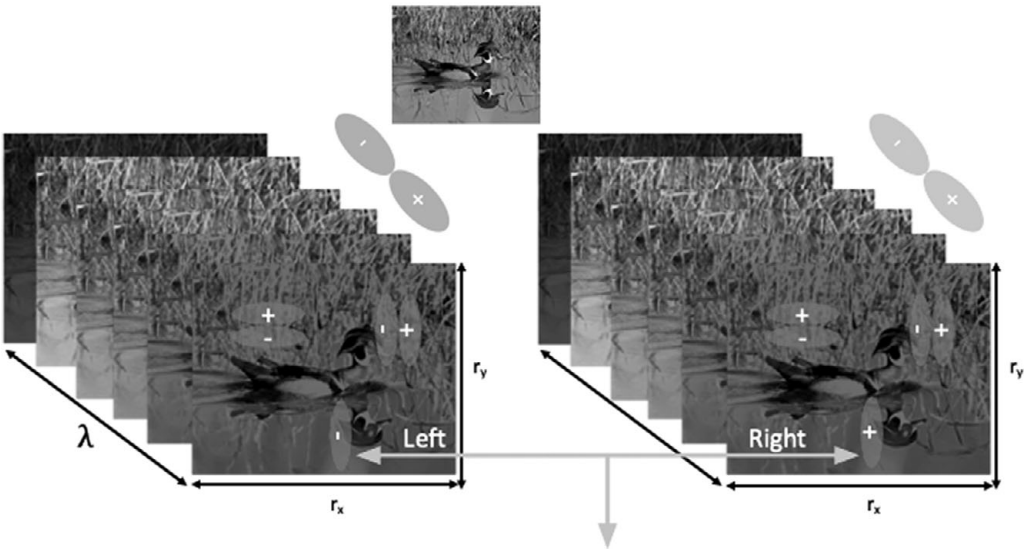


Figure 1.8 *Derivatives of the retinal image. A scene (top) is represented as spectral irradiance hypercubes for the left and right eye. The responses of neurons that compute the local differences, as indicated by several oval pairs with ±, approximate local partial derivatives. Differences can be taken across spatial location, across wavelength, across the spectral radiance measured by the two eyes, and across time (not shown). This figure should be viewed in color. The color version is available at https://color.psych.upenn.edu/supplements/earlyencoding/derivativesColorFig.pdf. The original color image was kindly provided by David Sparks.*

respect to spatial position, wavelength, and viewpoint (i.e., across the viewpoints provided by the left and right eyes). Receptive fields that respond to these derivatives include neurons that are pattern-selective (Priebe, 2016; Shapley & Lennie, 1985), cone-opponent (Shevell & Martin, 2017; Solomon & Lennie, 2007), and stereo disparity-selective (Cumming & DeAngelis, 2001). Partial derivatives with respect to time describe motion-selective neurons (Pasternak & Tadin, 2020; Wei, 2018).

The emphasis that Adelson and Bergen (1991) place on these derivatives is consistent with the generally accepted idea that it is the local change (contrast) in the spectral irradiance, not the absolute level of that irradiance, that provides the critical information used for perception (Shapley, 1986). Later in the chapter, we analyze psychophysical measurements of contrast sensitivity, which characterize quantitatively how small changes in spatial contrast are encoded by human vision.

An additional advantage of representations based on derivatives is that they are a highly compressible representation of naturally occurring spectral irradiance. The reason for this is that natural radiances tend to vary slowly, and thus many of the partial derivatives are near zero. A distribution with many repeated values may be compressed by coding the repeated values with tokens specified with a small number of bits, reserving tokens specified with a large number of bits for rarely occurring values (Cover & Thomas, 1991; Wandell, 1995).

1.5 Perceptual Inference

1.5.1 Ambiguity and Perceptual Processing

An important and consistent take-away from the analysis of sensory encoding is that the information available to the brain about the state of the external world is ambiguous: many different physical configurations produce the same sensory representation. A classic example is metamerism: there are only three classes of cone photoreceptors and different spectra produce identical triplets of responses in the L-, M-, and S-cones. Another well-known example is depth reconstruction: the three spatial dimensions of the light field are projected onto a two-dimensional retina, and many 3D shapes produce the same retinal image. Such many-to-one mappings are a reason why Helmholtz (1866, 1896) emphasized perceptual inference: the brain decodes the sensory representation to produce perceptions that are a likely guess about the state of the external world. Perception is an unconscious inference.

1.5.2 Mathematical Principles of Inference

The mathematical formulation of perceptual inference can be developed within a Bayesian probabilistic framework. Suppose $\mathbf{x}$ is a vector that describes some aspect of a scene. The entries of $\mathbf{x}$ might represent the spectral power density of a light entering the eye at a set of discretely sampled wavelengths, the pixel values of a displayed stimulus image, the optical flow vectors corresponding to a

viewed dynamic scene, or a full 3D scene description input to a computer graphics package. Now, suppose $\mathbf{y}$ is the sensory representation at some stage of the visual system produced when an observer views the scene described by $\mathbf{x}$. The entries of $\mathbf{y}$ might describe the retinal image, the excitations of each cone in the retinal mosaic, or the action potentials in a class of retinal ganglion cells.

Because sensory measurements are noisy, the relation between $\mathbf{y}$ and $\mathbf{x}$ is described by a conditional probability distribution, $p(\mathbf{y}|\mathbf{x})$. This distribution is referred to as the likelihood function. The likelihood can be thought of as a forward model that relates the scene parameters $\mathbf{x}$ to the sensory representation $\mathbf{y}$.

Within the Bayesian framework, the perceptual representation results from a choice the brain makes about the most likely scene given the observed sensory representation. Indeed, we can reverse the likelihood function, $p(\mathbf{y}|\mathbf{x})$, to obtain a conditional probability distribution $p(\mathbf{x}|\mathbf{y})$, which is called the posterior distribution. The posterior defines which are the more or less likely scenes, given the sensory measurements. To obtain the posterior, we use Bayes' rule (Bishop, 2006; Lee, 1989):

$$p(\mathbf{x}|\mathbf{y}) = K(\mathbf{y})p(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \tag{1.14}$$

where $K(\mathbf{y})$ is a normalizing factor that depends on $\mathbf{y}$ but not $\mathbf{x}$. This factor ensures that the posterior integrates to 1 for any value of $\mathbf{y}$. For many applications, our interest is in how the posterior depends on $\mathbf{x}$, and it is not necessary to compute $K(\mathbf{y})$.

Critically, $p(\mathbf{x})$ is a prior distribution that describes the statistical regularities of the scenes; how likely it is *a priori* that the world is in the state $\mathbf{x}$. A prior is essential because many scenes might have produced the same sensory measurements. Bayes' rule specifies how to combine the prior with the likelihood. Sometimes little is known about the prior. In these cases, using the Bayesian formulation directs our attention to learn more about it. The Bayesian formulation also forces us to make the forward model explicit in the form of the likelihood.

The posterior is a distribution over possible $\mathbf{x}$. We need a means of selecting a specific value, say $\hat{\mathbf{x}}$, to generate the percept. One common way to make a choice is to select a value $\hat{\mathbf{x}}$ that is most likely: the maximum *a posteriori* (MAP) estimate. Other possibilities, such as the mean of the posterior, are also commonly used. The interested reader is referred to the literature on Bayesian decision theory for more on this topic (e.g., Berger, 1985).

It is helpful to consider a simple example. Above we explained that the mean cone excitations at a location are a linear function of the radiance of a displayed image. Suppose we treat the spectral radiance on a display as the state of the world $\mathbf{x}$, with the entries of $\mathbf{x}$ appropriately ordered, and we denote the noisy cone excitations as $\mathbf{y}$. Then

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \boldsymbol{\epsilon} \tag{1.15}$$

for an appropriately arranged matrix $\mathbf{C}$, and where the noise in cone excitations is represented by the random variable $\boldsymbol{\epsilon}$. If we approximate $\boldsymbol{\epsilon}$ with a signal-independent zero-mean Gaussian distribution, we have

$$p(\mathbf{y}|\mathbf{x}) = norm(\mathbf{C}\mathbf{x}, \sigma_y^2 \mathbf{I}_y), \tag{1.16}$$

where *norm*() denotes the multivariate Gaussian distribution, σ_y^2 is the variance of the noise added to each mean cone excitation under the Gaussian approximation to the Poisson noise. The symbol $\mathbf{I}_y$ denotes the identity matrix with the same dimensionality as the vector $\mathbf{y}$.

We can also use a Gaussian distribution to describe a prior over $\mathbf{x}$:

$$p(\mathbf{x}) = norm(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \tag{1.17}$$

where the vector $\boldsymbol{\mu}_x$ and matrix $\boldsymbol{\Sigma}_x$ represent the mean and covariance of the prior.

Given the Gaussian likelihood and prior, the posterior is also Gaussian; its mean and covariance matrix may be computed analytically from the mean and covariance matrices of the likelihood and prior. This result follows from a standard identity that the product of two multivariate Gaussian distibutions is also a multivariate Gaussian (see Rasmussen & Williams, 2006; Brainard, 1995 provides the derivation in the context of the Bayesian posterior). In the case where the posterior is a multivariate Gaussian, its mean $\boldsymbol{\mu}_{x|y}$ provides the estimate of $\mathbf{x}$ that corresponds to both the posterior mean and the MAP estimate.

Figure 1.9 illustrates the idea for a simple example case. Suppose that the display has only two pixels and emits at only one wavelength. Then $\mathbf{x} = [x_1, x_2]^T$. We will assume that the radiance at each pixel of the display can range between 0 and 1. For natural images, there is a strong correlation between the radiance at neighboring pixels at the same wavelength (Burton & Moorehead, 1987; Tkacik *et al.*, 2011). A bivariate Gaussian prior distribution with this property is illustrated in the left panel of Figure 1.9. The mean of the prior is $\mathbf{x} = [0.5, 0.5]^T$ while the covariance matrix $\boldsymbol{\Sigma}_x$ corresponds to a common standard deviation of 0.127 and a correlation across the two pixels of 0.89. The strong correlation in the prior restricts the best guesses about the values of $\mathbf{x}$ relative to the full available range.

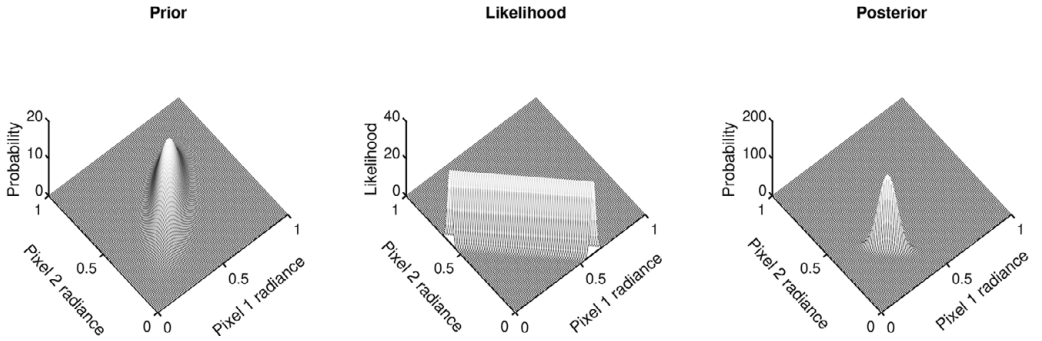


Figure 1.9 *Bayes' reconstruction. See description in text. For the prior and posterior, probability is given as the probability mass for a region of size* 0.01^2 *in the pixel radiance plane. Matlab code to produce this figure is available at https://github.com/DavidBrainard/BrainardFigListings.git (subdirectory scripts/MathPsychChapter/FigLinBayesExample, script Example.m).*

To compute a likelihood we need to know the nature of the sensory measurements. We suppose that there is just one cone and that it is equally sensitive to the radiance at the two display pixels. This gives us $\mathbf{C} = [0.5, 0.5]$. We assume that the mean excitation of the cone is perturbed by zero-mean Gaussian noise with standard deviation $\sigma_y = 0.01$. The middle panel of Figure 1.9 illustrates the likelihood for the specific cone excitation $\mathbf{y} = 0.3$: the likelihood $p(\mathbf{y} = 0.3|\mathbf{x})$ is plotted as a function of x_1 and x_2. This likelihood is highest along the ridge where the weighted sum of the pixel radiances sums to the observed cone excitation of 0.3. The likelihood falls off away from this ridge, with the rate of falloff determined by the magnitude of the noise. If the noise were smaller, the falloff would be faster and the likelihood ridge thinner, and conversely if the noise were larger, the falloff would be slower and the likelihood ridge wider. The likelihood alone tells us that $\mathbf{x}$ is unlikely to lie far from the ridge. At the same time, the likelihood makes explicit the ambiguity about $\mathbf{x}$ remaining after observing $\mathbf{y}$, with many values of $\mathbf{x}$ equally likely.

Bayes' rule specifies that the prior and likelihood should be combined using point-by-point multiplication over the pixel radiance plane [Equation (1.14)], and then normalized to form the posterior. The right panel of Figure 1.9 illustrates the result of this multiplication. The same result may be obtained directly by application of the analytic formulas for the posterior.

The posterior makes intuitive sense: it is large where both the prior and likelihood are large, and the resulting distribution is more concentrated than either the prior or likelihood alone. Although there is still uncertainty remaining in the posterior, it captures what we know about the scene when we combine the statistical regularities of the displayed images with the sensory measurement provided by the cone excitation.

1.5.3 Thresholds and Ideal Observer Theory

In this and the next sections, we show how ideas of perceptual inference as implemented through Bayes' rule help us understand perceptual processing. We begin with analysis of threshold measurements. A threshold is the minimum difference required for an observer to correctly discriminate between two stimuli; threshold measurements are a fundamental psychophysical tool. They are used to characterize perceptual performance and guide inferences about the neural mechanisms underlying this performance.

Consider, for example, discrimination between a uniform field and a contrast grating (see Figure 1.10). In a typical experiment, the observer is shown the uniform field and the grating in sequence, with the order randomized on each trial. The observer's task is to indicate which was presented first. In the experiment the stimulus contrast is titrated to a level at which the observer is correct, say, 80% of the time. The estimated contrast is the threshold.

Threshold measurements quantify the information needed by the visual system to make a basic perceptual decision: namely, that two stimuli differ. They involve small perturbations of the visual stimulus, and they may be thought of as assessing

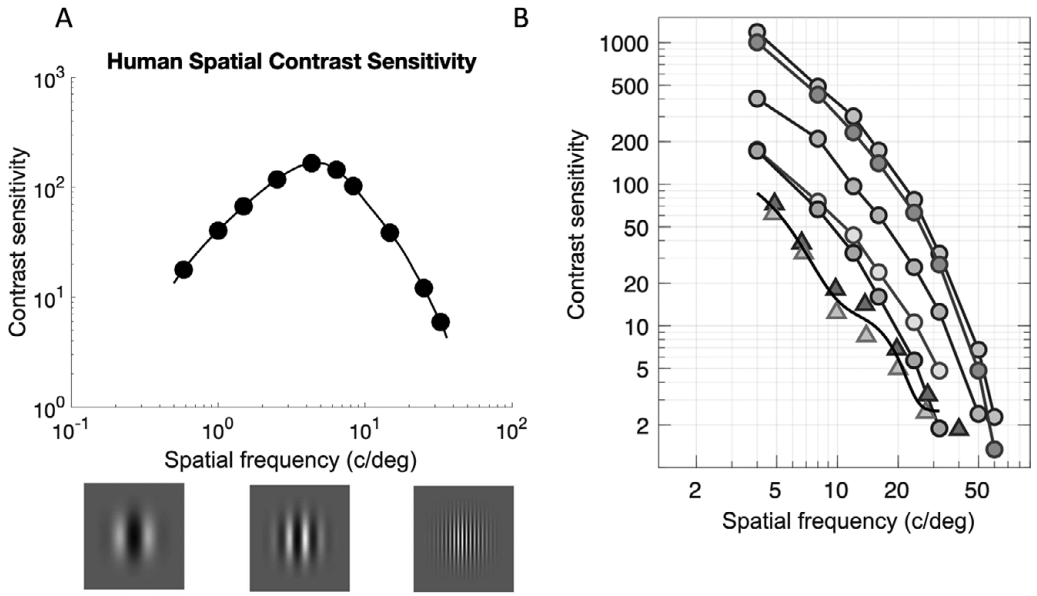


Figure 1.10 *Modeling the human contrast sensitivity function. (A) Sensitivity, defined as the inverse of threshold contrast, is plotted as a function of spatial frequency. The stimuli were small, equal-sized patches of contrast gratings. Replotted from De Valois, Morgan, and Snodderly (1974). The smooth curve replots the smooth curve in the original figure, while the solid points show the spatial frequencies on the smooth curve at which contrast sensitivity was measured. See the original figure for the actual sensitivity measurements through which the smooth curve was drawn. The thumbnails below the plot illustrate contrast grating patches at different spatial frequencies, but are not otherwise matched to the spatial frequency of the plot. (B) Triangles and black line: Human contrast sensitivity function for two observers, data from Banks, Geisler, and Bennett (1987). Grey circles/line: Contrast sensitivity of an ideal observer implemented at the level of the Poisson limited cone excitations, from Banks, Geisler, and Bennett (1987). Red circles/line: Ideal observer CSF with recent estimates of optics and mosaic properties. Blue circles/line: Computational observer CSF with decision rule determined using supervised machine learning. Green circles/line: Computational observer CSF additionally accounting for fixational drift. Purple circles/line: Computational observer CSF additionally incorporating a model of the transformation from excitations to photocurrent. This figure should be viewed in color. The color version is available at https://color.psych.upenn.edu/supplements/earlyencoding/csfColorFig.pdf. If you are nonetheless viewing a grayscale version of the figure, the order of the colors of the ideal/computational observer CSFs from top to bottom is: gray, red, blue, green, purple. After Figure 6 of Cottaris* et al. *(2020).*

sensitivity to derivatives of the retinal image. In this way, thresholds are connected to the ideas introduced above about the importance of derivatives of the spectral radiance as a basis for visual processing.

Figure 1.10 shows the threshold for contrast gratings measured as a function of grating spatial frequency: this is called the spatial contrast sensitivity function (CSF). When measured with static or very slowly moving gratings, the human CSF has an inverted U-shape: the highest contrast sensitivity is between three and six cycles per degree, with lower sensitivity at higher and lower spatial frequencies. Because any image may be synthesized by a weighted superposition of sinusoidal gratings (Bracewell, 1978), the CSF characterizes the sensitivity to basic stimulus components. Because the visual system as a whole is neither shift-invariant nor linear, however, the CSF is a useful but incomplete description of sensitivity.

We would like to understand how the human contrast sensitivity function is limited by the properties of the visual components described in this chapter. Bayes' rule provides a way to build this understanding by linking the initial encoding to performance on the psychophysical threshold detection task. Analyses of this sort are called ideal observer theory (Geisler, 1989). Ideal observer theory allows us to estimate the extent to which discrimination performance is limited by the early visual encoding. Relevant factors include blurring by the eye's optics, which reduces the retinal contrast of a grating stimulus, spatial sampling by the cone mosaic, and the Poisson variability in the cone excitations. Of particular interest is separating aspects of visual performance that are tightly coupled to these factors from aspects that are limited by processes not incorporated into the ideal observer calculation.

So, how do we use Bayes to predict performance in the two-interval forced choice task described above? We use the terms reference stimulus and comparison stimulus to describe the two stimuli being discriminated. In this example the reference stimulus is a spatially uniform field and the comparison stimulus is a patch of contrast grating with known spatial frequency, orientation, size, and contrast; but, the ideas we develop here apply to any two stimuli being discriminated.

Using the computational methods described in this chapter, we compute the mean cone excitations to the reference and comparison stimuli. Let $\mathbf{u}_r$ be the vector of mean cone mosaic excitations in response to the reference stimulus and let $\mathbf{u}_c$ be the vector of mean cone excitations in response to the comparison stimulus. In the two-interval forced choice task, the observer must indicate whether the reference came first followed by the comparison, or the other way around. We thus form two concatenated vectors, $\mathbf{u}_1 = [\mathbf{u}_r, \mathbf{u}_c]$ and $\mathbf{u}_2 = [\mathbf{u}_c, \mathbf{u}_r]$.

To apply Bayes' rule to this problem, we think of the scene as described by a binary random variable. This variable, x, can take on value 1 or 2. These values represent the reference first and reference second possibilities that can occur on each trial. The prior probability $p(x)$ is given by

$$p(x = 1) = 0.5;\ p(x = 2) = 0.5. \tag{1.18}$$

The data available to the observer to make a response of $x = 1$ or $x = 2$ are the pattern of observed cone excitations across the two intervals, which we will denote by $\mathbf{y}$. We know that for $x = 1$, each entry of $\mathbf{y}$ is an independent Poisson random

variable with mean given by the corresponding entry of $\mathbf{u}_1$, while for $x = 2$ the means are given by the corresponding entries of $\mathbf{u}_2$. From this, we have for the posterior:

$$p(x = 1|\mathbf{y}) = Kp(\mathbf{y}|x = 1)p(x = 1) = K\prod_i p(y_i|x = 1)p(x = 1), \tag{1.19}$$

where y_i denotes the ith entry of $\mathbf{y}$ and we have explicitly expressed the joint distribution of independent random variables as the product of their individual distributions. K is a normalizing constant whose value we need not calculate.

We substitute the expression for the probability mass function of a Poisson random variable and the value of $p(x = 1)$ to obtain

$$p(x = 1|\mathbf{y}) = K\prod_i \frac{{u_{1i}}^{y_i}e^{-u_{1i}}}{y_i!}0.5, \tag{1.20}$$

where u_{1i} denotes the ith entry of $\mathbf{u}_1$. Similarly, we have

$$p(x = 2|\mathbf{y}) = K\prod_i \frac{{u_{2i}}^{y_i}e^{-u_{2i}}}{y_i!}0.5. \tag{1.21}$$

To maximize the percent correct on the task, the observer should compare $p(x = 1|\mathbf{y})$ with $p(x = 2|\mathbf{y})$ and indicate 1 or 2 according to which is larger. It is instructive to implement this comparison in terms of the difference of the logs of $p(x = 1|\mathbf{y})$ and $p(x = 2|\mathbf{y})$, with a response of 1 corresponding to a difference greater than or equal to 0 and a response of 2 corresponding to a difference less than 0. Writing the difference of logs explicitly and simplifying, we have decision variable

$$\delta = \sum_i y_i \log\left(\frac{u_{1i}}{u_{2i}}\right) + \sum_i (u_{2i} - u_{1i}). \tag{1.22}$$

An observer who responds according to the sign of δ will maximize the percent correct. The value of the percent correct depends on how δ is distributed when $x = 1$ and $x = 2$. Geisler (1984) provides a Gaussian approximation to these distributions, which may be used to obtain the corresponding percent correct. As with the human psychophysical experiment, contrast may be titrated to find the ideal observer threshold contrast, that which leads to the ideal observer having the criterion percent correct.

Figure 1.10B shows the ideal observer contrast sensitivity for human foveal viewing (gray circles/line), along with psychophysical measurements of human contrast sensitivity at spatial frequencies increasing from 5 cpd, and with the measurements (triangles/black line) made with stimuli matched to those used in the ideal observer calculations (Banks, Geisler, & Bennett, 1987). As with the human data at higher spatial frequencies, the ideal observer contrast sensitivity function falls off as spatial frequency increases; the slope of this falloff closely resembles that of the human observer. This correspondence suggests that the factors that cause the human falloff share basic features with those included in

the ideal observer calculation. Here the primary factor is blur from the eye's optics and cone apertures, both of which reduce the contrast captured by spatial variation in the cone excitations.

The ideal observer CSF also differs from the human measurements. One difference is that the overall sensitivity of the ideal observer is markedly higher than that of the human observer. The Poisson noise in the cone excitations limits the ideal observer sensitivity. The fact that incorporating only this noise source leads to an ideal observer more sensitive than the human tells us that additional factors limit human sensitivity and motivates study of what these additional factors are.

One approach is to define a single "efficiency" parameter representing an omnibus loss of information by the actual visual system relative to an ideal observer calculation. This is often sufficient to bring ideal observer predictions into alignment with measured human performance (Burge, 2020), as is true in the case of the ideal and human CSF rolloff at high spatial frequencies. The efficiency parameter can be thought of as capturing the effect of additional noise in the human visual system, not included in the ideal observer calculation, whose effect on performance is stimulus-independent.

It is important to note, however, that the difference between ideal and human performance is not fully explained by a single efficiency parameter. For example, the ideal observer CSF does not roll off at low spatial frequencies but the human CSF does. The factors that produce the measured low-spatial-frequency rolloff are not included in the ideal observer calculations presented here. As with the difference in overall sensitivity, the difference between ideal and human CSF at low spatial frequencies motivates investigation of what additional factors in the human visual system account for the difference.

1.5.4 Computational Observers

The ideal observer calculation used by Banks, Geisler, & Bennett, (1987) employed a simplified model of the eye's point spread function and cone mosaic, and this simplification enabled efficient computation of ideal observer performance. In two recent papers, Cottaris *et al.* (2019, 2020) employed computational methods to examine the effect of more recent estimates of the point spread function (Thibos *et al.*, 2002) and a more detailed model of the foveal mosaic on performance. These had only a modest effect on the predictions (Figure 1.10B, red circles/line).

The ideal observer developed above has full knowledge of mean cone excitations and Poisson structure of the noise, so that the observer's performance is not degraded by stimulus uncertainty (Geisler, 2018; Pelli, 1985). Cottaris *et al.* (2019) relaxed this assumption by replacing the ideal observer decision rule with a decision rule based on a trained linear classifier (C. D. Manning, Raghavean, & Schutze, 2008; Schölkopf *et al.*, 2002). The classifier measured the match of the data to a template that had the same spatial structure as the stimuli. The decision boundary was optimized in the presence of noise. The need to partially learn the decision rule reduced the absolute level of ideal observer performance

while retaining the same CSF shape (blue circles/line in Figure 1.10B). Cottaris *et al*. (2020) then introduced a computational model of fixational eye movements (Mergenthaler & Engbert, 2007; see also Engbert & Kliegl, 2004) and showed that an approach to handling the stimulus motion blur introduced by these movements further reduced performance (green circles/line). Finally, Cottaris *et al*. (2020) introduced a computational model of the transformation from excitations to electrical photocurrent, which included both gain control and additional noise. Accounting for this transformation brought computational observer performance into approximate alignment with the human measurements at the higher spatial frequncies (purple circles/line).

This analysis outlines a set of factors that together provide an account of the high-spatial-frequency limb of the human spatial CSF, capturing both the shape and absolute level of this important measure of performance. For the purposes of the present chapter, we emphasize less the specific elements of the account, which will surely be refined by future research, but rather the way the mathematical principles are combined with computational modeling with the goal of accounting for the full richness of the visual system. The combination of principles and computations accounts for factors that are beyond what is possible using analytic calculations alone.

1.5.5 Image Reconstruction

The ideal observer and computational observer development above applies Bayesian inference to the analysis of threshold measurements. Thresholds characterize the limits of visual performance, and the analyses illustrate how threshold performance can be linked to quantitative measurements of physiological optics, retinal anatomy, and retinal physiology. Not all vision is threshold vision, however. Sometimes we are interested in predicting what clearly visible stimuli look like (e.g., "that apple looks red") or how similar easily distinguishable objects appear (e.g., "the color of the apple appears more similar to the color of the tomato than it does to the color of the banana"). There are a number of methods for studying suprathreshold vision. These include asymmetric matching (Brainard & Wandell, 1992; Burnham, Evans, & Newhall, 1957; Wandell, 1995) and various scaling techniques (T. F. Cox & Cox, 2001; Knoblauch & Maloney, 2012; Maloney & Yang, 2003). We will not treat these methods here. Below, however, we illustrate how Bayesian methods can be used to understand how the initial visual encoding shapes the perceptual inferences that can be made about suprathreshold stimuli.

In our introduction to Bayes' rule, we illustrated the core ideas by considering reconstruction of a two-pixel image from the excitations of a single cone, using both a Gaussian and a Gaussian likelihood. As computer power has increased, these same Bayesian principles have been applied to increasingly large perceptual problems. As we illustrate here, it is now possible to reconstruct an estimate of a full displayed color image from a realistic model of cone excitations using the Poisson likelihood (Zhang, Cottaris, & Brainard, 2021).

The forward computation starts with the displayed image $\mathbf{x}$ and computes the cone excitations $\mathbf{y}$. The vector $\mathbf{x}$ can be thought of as the concatenation of the linearized and rasterized pixel values for each of the red, green, and blue channels of the display. Using the Poisson noise model of the cone excitations, we compute the likelihood of observed cone excitations $p(\mathbf{y}|\mathbf{x})$. Here the vector $\mathbf{y}$ is simply a list of the excitations of each cone in the mosaic. Because the mean cone excitations are a linear function of the display pixel values, we can write for these mean excitations

$$\bar{y} = \mathbf{R}\mathbf{x} \tag{1.23}$$

for some matrix $\mathbf{R}$. Each column of this matrix may be computed as the vector of cone excitations produced when one pixel is at its maximum value for one color channel, with the display values for all other pixels and color channels set to zero, and these computations may be implemented in software such as ISETBio to determine explicitly the matrix $\mathbf{R}$ (Zhang, Cottaris, & Brainard, 2021). This yields for the likelihood

$$p(\mathbf{y}|\mathbf{x}) = Poisson(\mathbf{R}\mathbf{x}), \tag{1.24}$$

where *Poisson*() denotes the result of Poisson noise applied independently to its vector argument by taking each entry of the argument as the corresponding Poisson mean.

Next, we specify a prior distribution $p(\mathbf{x})$ for natural images. Natural images have a great deal of structure (Simoncelli, 2005), and a full statistical description of this structure is not currently available. There are two robust regularities of natural images, however, that can be described by a multivariate Gaussian. The first is that within a single wavelength band, the spectral radiances at nearby image locations are highly correlated (Field, 1987; Pratt, 1978; Ruderman, Cronin, & Chiao, 1998). The second regularity is that at a single position, values in nearby wavelength bands are highly correlated (Burton & Moorehead, 1987; Tkacik *et al.*, 2011). This is a consequence of the relatively smooth spectral functions one observes in nature (Cohen, 1964; Maloney, 1986; Vrhel, Gershon, & Iwan, 1994). These two observations may be used to construct a covariance matrix for a multivariate Gaussian that describes the second-order statistics of natural images. Together with the average image, these provide a Gaussian image prior.

With the likelihood and prior, we can construct an estimate of the image given a vector of cone excitations. As with many calculations described in this chapter, the principles of Bayesian estimation guide the way, but once we introduce the Poisson likelihood, we turned to numerical computational methods to find the solution.

We used ISETBio to reconstruct images from cone excitations, with the Poisson likelihood and Gaussian image prior described above. We reconstructed images for retinal patches at various visual field eccentricities. As visual field eccentricity increases, the point spread of the retinal image becomes more blurred and the density with which the cones sample the image decreases (Figures 1.5, 1.6, and 1.7). Thus, less information becomes available to the visual system in the

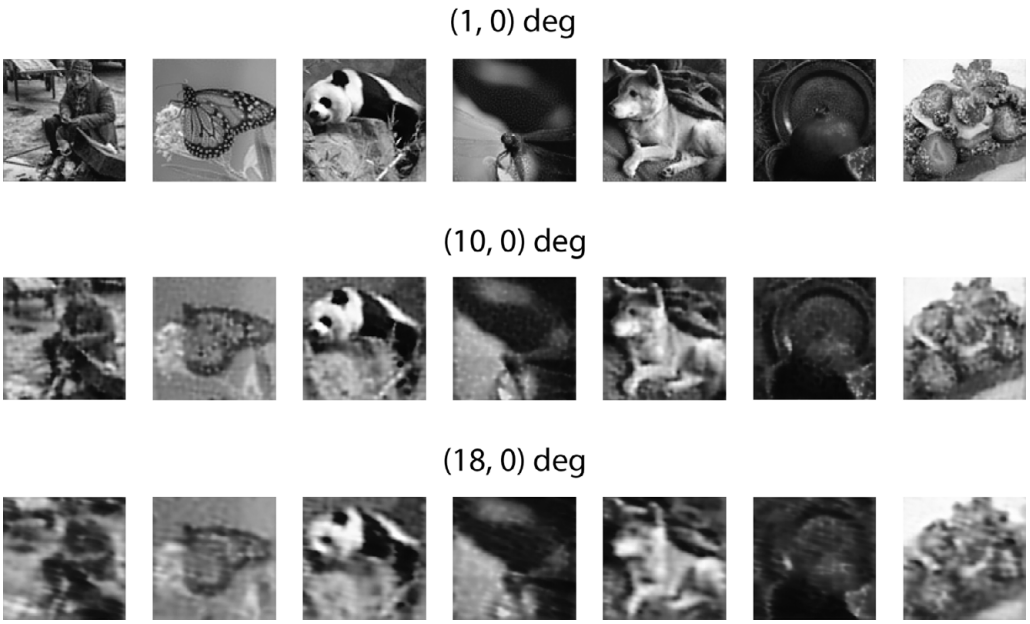


Figure 1.11 *Image reconstructions from cone excitations at three retinal eccentricities. Each row shows reconstructions of seven images using the Bayesian method and Poisson likelihood and multivariate Gaussian prior. The reconstructions at 1° eccentricity are close to veridical, with increasing distortions seen at the 10 and 18° locations. Each original and reconstructed image was represented at a pixel resolution of 128 × 128, and the extent of each image on the retina was* $1^\circ \times 1^\circ$. *The mean excitation of the cones was* 10^5 *excitations per cone, so the simulation corresponds to a relatively high signal-to-noise regime. The parameters of the Gaussian prior were fit to 16 × 16 pixel patches of images from the ImageNet ILSVRC data set (www.image-net.org), and extended in an overlapping blockwise fashion to the higher image pixel resolution. This figure should be viewed in color. The color version is available at https://color.psych.upenn.edu/supplements/earlyencoding/GaussianReconColorFig.pdf. Figure courtesy Lingqi Zhang. See Zhang, Cottaris, and Brainard (2021) for a more extended discussion of Bayesian image reconstruction and the general methods used to produce this figure.*

peripheral visual field. The effect of this loss for reconstruction depends on the prior. Although the information loss means that two images whose cone excitations are different in the fovea can produce the same cone excitations in the periphery, this ambiguity need not degrade the reconstructions if the probability that one of the two images will occur is small.

The reconstructions in Figure 1.11 show the effect of information loss at the level of the cone excitations, in the context of the Gaussian image prior. The reconstructed image quality in the periphery is worse than in the fovea, but many objects remain recognizable from the peripheral reconstructions. Moreover, there

are interesting interactions between the likelihood and prior. For example, the recovery of color can be better in the fovea and more peripheral locations than it is in the mid-periphery (see images of strawberries in Figure 1.11, for example). Zhang, Cottaris, and Brainard (2021) describe the reconstruction approach to analyzing the initial visual encoding in more detail, extending the ideas to a more realistic prior than the Gaussian, and showing a number of calculations that use image reconstruction to examine how prior and likelihood interact to support both color and spatial vision (see also Brainard, Williams, & Hofer, 2008).

Image reconstruction computations provide useful insights about how statistical regularities in natural scenes interact with the sensory measurements to guide perception. But, it is important to bear in mind that reconstruction of displayed images is not the task for which visual perception evolved. Rather, we view the task of perception to reconstruct the properties and positions of objects in the three-dimensional environment. The Bayesian ideas presented here have applicability to this task as well (Knill & Richards, 1996), but a computational solution that is as effective as human vision currently remains elusive. This is an area where recent progress in machine learning and deep neural networks may provide new insights.

1.5.6 Optimizing Sensory Measurements

Earlier in this chapter, we explained that the visual system appears to extract information about motion, color, and pattern from the pattern of cone excitations by estimating the local derivatives of various quantities (Adelson & Bergen, 1991). The Bayesian framework provides a quantitative framework for addressing how to optimize which signals should be transduced by a sensory system when the goal is reconstruction of the state of the environment, as well as how the sensory signals should be summarized (e.g., in the form of local derivatives) for further processing. Indeed, the Bayesian image reconstruction methods developed here point towards the ingredients required for a full analysis of such questions. To know what measurements we should make, we first need to know the prior distribution over the environmental states that an organism will encounter. We then need a parameterized set of candidate likelihood functions, each of which describes a feasible arrangement of the sensory apparatus and (if desired) associated early processing. This information allows us to compute the posterior over the environmental states for any candidate likelihood function, and we can ask how well different sensory measurements constrain the posterior, averaging this information over the environmental states described by the prior. Developing a parameterized set of candidate likelihood functions requires an understanding of what biological constraints apply to the sensory system. Also required is an understanding of the cost of different types of error in the resultant perceptual representation (the loss function; Berger, 1985), as well as how the cost of error should be balanced against the energetic cost of making and processing the sensory measurements (Balasubramanian, Kimber, & Berry, 2001; Koch *et al.*, 2004; Laughlin, 2001). A number of authors have pursued questions of optimizing sensory measurements in this manner (Garrigan *et al.*, 2010; Levin, Durand, &

Freeman, 2008; J. R. Manning & Brainard, 2009; Zhang, Cottaris, & Brainard, 2021).[10] It would be interesting to compare the results of an analysis of this sort to the Adelson/Bergen conjecture that approximations to local derivatives represent an optimal measurement set.

1.6 Summary and Conclusions

To focus on the mathematics of the initial visual encoding, we introduce vision science from the point of view of a forward calculation: physics of the stimulus, image formation, and quantitative system modeling. The key mathematical principles are linear algebra, shift-invariant linear systems, and specification of sensory noise. The mathematics of vision science shares much in common with the mathematics of many fields of science and engineering.

After expressing and implementing the forward calculations, we explore the mathematics of Helmholtz's hypothesis: people perceive a stimulus that is the most likely explanation of the cone excitations. We use Bayesian inference methods to clarify the uncertainty about the encoded signal. This approach requires that we confront the problem of establishing priors on the signal. There is a close connection between Helmholtz's unconscious inference and Bayesian inference; the latter may be thought of as a quantitative implementation of Helmholtz's idea.

The approach we describe has a long and accomplished tradition. But, it is not the only valid way to make progress in vision science; several other approaches are important. A quantitative study of behavioral rules can be very informative. For example, color appearance matching was a largely behavioral exploration at first; an understanding of the physics of the signal and the biological underpinnings followed later. Also, neurobiological measures can be helpful. Anatomical and functional measurements that characterize the properties of multiple pathways within the visual system – including multiple types of retinal ganglion cells and multiple pathways through the visual cortex – are useful guides to understanding visual specializations and computations, particularly for stages of vision beyond the initial encoding. Finally, engineering work to build functional artificial visual systems continues to be very helpful in understanding vision: a classic principle states that the best way to demonstrate you understand a system is to build one that does the same thing. Engineering efforts continue to clarify features that we might look for in the nervous system, as well as why certain behavioral patterns emerge.

The field of vision science is large and vigorous enough that there is no need to choose a single approach. We are inspired by the fact that different investigators adopt different approaches, all seeking to gain understanding. To the student thinking about how to approach vision science, we offer advice from an American philosopher who commented about making difficult decisions: "When you come to a fork in the road, take it" (Yogi Berra).

10 The formalism used in these analyses is interestingly similar to that underlying Bayesian adaptive psychophysical procedures (Watson, 2017; Watson & Pelli, 1983).

1.7 Related Literature

This chapter introduces key mathematical and computational approaches to understanding the initial visual encoding. A number of the mathematical ideas we present here are developed in more detail by Wandell (1995), and the classic treatment of visual perception by Cornsweet (1970) remains a valuable introduction to the field, as does Rodieck (1998). Principles of ray tracing are introduced in many computer graphics texts (e.g., Pharr, Jakob, & Humphreys, 2016); similarly many texts introduce optics (e.g., Hecht, 2017). In the context of the retinal image and cone excitations specifically, Packer and Williams (2003), Pugh (1988), and Yellott, Wandell, and Cornsweet (1984) are useful. Brainard and Stockman (2010) elaborate in more detail on using linear algebra in support of colorimetric applications. Although we do not treat the Fourier transform and frequency domain representations in this chapter, the reader who wishes to specialize in this field will want to learn about these ideas. Two useful sources are Bracewell (1978) and Pratt (1978). Useful introductions to statistical inference include Bishop (2006) and Duda, Hart, and Stork (2001).

Acknowledgments

We thank Nicolas Cottaris and Lingqi Zhang for providing figures for this chapter. We also thank Amy Ni, Nicolas Cottaris, Lingqi Zhang, Joyce Farrell, Eline Kupers, Heiko Schutt, and Greg Ashby for useful comments on the manuscript.

References

Adelson, E. H., & Bergen, J. (1991). The plenoptic function and the elements of early vision. In M. Landy & J. Movshon (Eds.), *Computational models of visual processing* (pp. 3–20). Cambridge, MA: MIT Press.

Adelson, E. H., & Wang, J. Y. A. (1992, February). Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 99–106.

Ayscough, J. (1755). *Light field image.* https://commons.wikimedia.org/wiki/File:1755_james_ayscough.jpg. (Accessed: July 24, 2021.)

Balasubramanian, V., Kimber, D., & Berry, M. J. (2001). Metabolically efficient information processing. *Neural Computation*, *13*(4), 799–815.

Banks, M. S., Geisler, W. S., & Bennett, P. J. (1987). The physical limits of grating visibility. *Vision Research*, *27*(11), 1915–1924.

Baylor, D. A., Lamb, T. D., & Yau, K. W. (1979). Responses of retinal rods to single photons. *Journal of Physiology*, *288*(Mar), 613–634.

Berger, T. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer-Verlag.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer Science + Business Media LLC.

Bracewell, R. (1978). *The Fourier transform and its applications*. New York: McGraw-Hill.

Brainard, D. H. (1995). *An ideal observer for appearance: Reconstruction from samples* (Tech. Rep. No. 95-1). UCSB Vision Labs Technical Report.

Brainard, D. H. (2015). Color and the cone mosaic. *Annual Review of Vision Science*, *1*, 519–546.

Brainard, D. H., & Stockman, A. (2010). Colorimetry. In M. Bass et al. (Eds.), *The Optical Society of America handbook of optics, 3rd edition, Volume III: Vision and vision optics* (pp. 10.1–10.56). New York: McGraw-Hill.

Brainard, D. H., & Wandell, B. A. (1992). Asymmetric color-matching: How color appearance depends on the illuminant. *Journal of the Optical Society of America A*, *9*(9), 1433–1448.

Brainard, D. H., Williams, D. R., & Hofer, H. (2008). Trichromatic reconstruction from the interleaved cone mosaic: Bayesian model and the color appearance of small spots. *Journal of Vision*, *8*(5), 1–23.

Branwyn, G. (2016, September). *Sky angles.* https://makezine.com/2016/09/16/measuring-tip-ruler/sky_angles/. (Accessed: August 8, 2021.)

Burge, J. (2020). Image-computable ideal observers for tasks with natural stimuli. *Annual Review of Vision Science*, *6*, 491–517.

Burnham, R., Evans, R., & Newhall, S. (1957). Prediction of color appearance with different adaptation illuminations. *Journal of the Optical Society of America*, *47*(1), 35–42.

Burns, S. A., Elsner, A. E., Lobes, L. A., Jr, & Doft, B. H. (1987, April). A psychophysical technique for measuring cone photopigment bleaching. *Investigative Ophthalmology & Visual Science*, *28*(4), 711–717.

Burton, G. J., & Moorehead, I. R. (1987). Color and spatial structure in natural images. *Applied Optics*, *26*(1), 157–170.

Cahan, D. (1993). *Hermann von Helmholtz and the foundations of nineteenth-century science*. Berkeley, CA: University of California Press.

Canon U.S.A., Inc. (2017, April). *Introduction to dual pixel autofocus.* www.usa.canon.com/internet/portal/us/home/learn/education/topics/article/2018/July/Intro-to-Dual-Pixel-Autofocus-(DPAF)/Intro-to-Dual-Pixel-Autofocus-(DPAF). (Accessed: July 8, 2021.)

CIE (1986). *Colorimetry, second edition* (Report No. 15.2). Vienna: Bureau Central de la CIE.

CIE (2007). *Fundamental chromaticity diagram with physiological axes - parts 1 and 2. technical report 170-1*. Vienna: Bureau Central de la CIE.

Cohen, J. (1964). Dependency of the spectral reflectance curves of the Munsell color chips. *Psychonomic Science*, *1*, 369–370.

Cornsweet, T. (1970). *Visual perception*. New York: Academic Press.

Cottaris, N. P., Jiang, H., Ding, X., Wandell, B. A., & Brainard, D. H. (2019). A computational observer model of spatial contrast sensitivity: Effects of wavefront-based optics, cone mosaic structure, and inference engine. *Journal of Vision*, *19*(4), 8.

Cottaris, N. P., Wandell, B. A., Rieke, F., & Brainard, D. H. (2020). A computational observer model of spatial contrast sensitivity: Effects of photocurrent encoding, fixational eye movements, and inference engine. *Journal of Vision*, *20*(7), 17.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons.

Cox, D. D., & Dean, T. (2014, September). Neural networks and neuroscience-inspired computer vision. *Current Biology*, *24*(18), R921–R929.

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Cumming, B. G., & DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual Review of Neuroscience*, *24*(1), 203–238.

Curcio, C., Sloan, K., Kalina, R., & Hendrickson, A. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, *292*(4), 497–523.

Da Vinci, L. (1970). *The notebooks of Leonardo da Vinci* (Vol. 1; J. P. Richter, Ed.). New York: Dover.

De Valois, R. L., Morgan, H., & Snodderly, D. M. (1974). Psychophysical studies of monkey vision–III. Spatial luminance contrast sensitivity tests of macaque and human observers. *Vision Research*, *14*(1), 75–81.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification and scene analysis,* 2nd ed. New York: John Wiley & Sons.

Engbert, R., & Kliegl, R. (2004). Microsaccades keep the eyes' balance during fixation. *Psychological Science*, *15*(6), 431–436.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Gamlin, P. D. R., McDougal, D. H., Pokorny, J., Smith, V. C., Yau, K. W., & Dacey, D. M. (2007). Human and macaque pupil responses driven by melanopsin-containing retinal ganglion cells. *Vision Research*, *47*(7), 946–954.

Garrigan, P., Ratliff, C. P., Klein, J. M., Sterling, P., Brainard, D. H., & Balasubramanian, V. (2010). Design of a trichromatic cone array. *PLoS Computational Biology*, *6*(2), e1000677.

Geisler, W. S. (1984). Physical limits of acuity and hyperacuity. *Journal of the Optical Society of America A*, *1*(7), 775–782.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*(2), 267–314.

Geisler, W. S. (2018). Psychometric functions of uncertain template matching observers. *Journal of Vision*, *18*(2), 1.

Gershun, A. (1939). The light field. *Journal of Mathematical Physics*, *18*(1–4), 51–151.

Grassmann, H. (1853). Zur Theorie der Farbenmischung. *Annalen der Physik und Chemie*, *165*, 69–84.

Hattar, S., Liao, H. W., Takao, M., Berson, D. M., & Yau, K. W. (2002). Melanopsin-containing retinal ganglion cells: Architecture, projections, and intrinsic photosensitivity. *Science*, *295*(5557), 1065–1070.

Hecht, E. (2017). *Optics*, 5th ed. Boston, MA: Pearson.

Hecht, S., Schlaer, S., & Pirenne, M. (1942). Energy, quanta and vision. *Journal of the Optical Society of America*, *38*(6), 196–208.

Helmholtz, H. (1866). *Handbuch der physiologischen Optik II* (3rd ed., 1911) (pp. 243–244). Hamburg: Voss.

Helmholtz, H. (1896). *Physiological optics*. New York: Dover.

Hofer, H., & Williams, D. R. (2014). Color vision and the retinal mosaic. In L. M. Chalupa & J. S. Werner (Eds.), *The new visual neurosciences* (pp. 469–483). Cambridge, MA: MIT Press.

Hunt, R. W. G. (2004). *The reproduction of colour*, 6th ed. Chichester: John Wiley & Sons.

Jaeken, B., & Artal, P. (2012). Optical quality of emmetropic and myopic eyes in the periphery measured with high-angular resolution. *Investigative Ophthalmology and Visual Science*, *53*(7), 3405–3413.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R (use R!)*. New York: Springer-Verlag.

Koch, K., McLean, J., Berry, M., Sterling, P., Balasubramanian, V., & Freed, M. A. (2004). Efficiency of information transmission by retinal ganglion cells. *Current Biology*, *14*(17), 1523–1530.

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, *11*(4), 475–480.

Lee, P. M. (1989). *Bayesian statistics*. London: Oxford University Press.

Levin, A., Durand, F., & Freeman, W. T. (2008). *Understanding camera trade-offs through a Bayesian analysis of light field projections* (Report). Cambridge, MA: MIT Press.

Lian, T., MacKenzie, K. J., Brainard, D. H., Cottaris, N. P., & Wandell, B. A. (2019). Ray tracing 3D spectral scenes through human optics models. *Journal of Vision*, *19*(12), 23.

Maloney, L. T. (1986). Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *Journal of the Optical Society of America A*, *3*(10), 1673–1683.

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8), 573–585.

Manning, C. D., Raghavean, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Manning, J. R., & Brainard, D. H. (2009). Optimal design of photoreceptor mosaics: Why we do not see color at night. *Visual Neuroscience*, *26*(1), 5–19.

Marimont, D. H., & Wandell, B. A. (1994, December). Matching color images: The effects of axial chromatic aberration. *Journal of the Optical Society of America A*, *11*(12), 3113–3122.

Maxwell, J. (1860). On the theory of compound colours and the relations of the colours of the spectrum. *Philosophical Transactions of the Royal Society of London*, *150*, 57–84.

Mergenthaler, K., & Engbert, R. (2007). Modeling the control of fixational eye movements with neurophysiological delays. *Physical Review Letters*, *98*(13), 138104.

Mlinar, M. (2016, May). *Image processing methods for image sensors with phase detection pixels* (No. 9338380).

Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., & Hanrahan, P. (2005). *Light field photography with a hand-held plenoptic camera* (Unpublished doctoral dissertation, Stanford University).

Packer, O., & Williams, D. R. (2003). Light, the retinal image, and photoreceptors. In S. K. Shevell (Ed.), *The science of color,* 2nd ed. (pp. 41–102). Oxford: Optical Society of America/Elsevier Ltd.

Pasternak, T., & Tadin, D. (2020). Linking neuronal direction selectivity to perceptual decisions about visual motion. *Annual Review of Vision Science*, *6*, 335–362.

Pelli, D. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*, *2*(9), 1508–1532.

Pharr, M., Jakob, W., & Humphreys, G. (2016). *Physically based rendering: From theory to implementation*. San Francisco, CA: Morgan Kaufmann.

Polans, J., Jaeken, B., McNabb, R. P., Artal, P., & Izatt, J. A. (2015). Wide-field optical model of the human eye with asymmetrically tilted and decentered lens that reproduces measured ocular aberrations. *Optica*, *2*(2), 124–134.

Pratt, W. K. (1978). *Digital image processing*. New York: John Wiley & Sons.

Priebe, N. J. (2016). Mechanisms of orientation selectivity in the primary visual cortex. *Annual Review of Vision Science*, *2*, 85–107.

Pugh, J. (1988). Vision: Physics and retinal physiology. In R. Atkinson, R. Herrnstein, G. Lindsey, & R. Luce (Eds.), *Stevens' handbook of experimental psychology*, 2nd ed. (Vol. 1, pp. 75–163). New York: John Wiley & Sons.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Rodieck, R. (1998). *The first steps in seeing*. Sunderland, MA: Sinauer.

Ruderman, D. L., Cronin, T. W., & Chiao, C. C. (1998). Statistics of cone responses to natural images: Implications for visual coding. *Journal of the Optical Society of America A*, *15*(8), 2036–2045.

Rushton, W. (1972). Pigments and signals in colour vision. *Journal of Physiology*, *220*(3), 1P–31P.

Schölkopf, B., Smola, A. J., Bach, F., *et al.* (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Shapley, R. M. (1986). The importance of contrast for the activity of single neurons, the VEP and perception. *Vision Research*, *26*(1), 45–62.

Shapley, R. M., & Lennie, P. (1985). Spatial frequency analysis in the visual system. *Annual Review of Neuroscience*, *8*(1), 547–581.

Shevell, S. K., & Martin, P. R. (2017). Color opponency: Tutorial. *Journal of the Optical Society of America A*, *34*(7), 1099–1108.

Simoncelli, E. P. (2005). Statistical modeling of photographic images. In A. Bovik (Ed.), *Handbook of image and video processing* (pp. 431–441). New York: Academic Press.

Solomon, S., & Lennie, P. (2007). The machinery of colour vision. *Nature Reviews Neuroscience*, *8*(4), 276–286.

Stiles, W., & Burch, J. (1959). NPL colour-matching investigation: Final report (1958). *Optica Acta*, *6*, 1–26.

Stockman, A., & Sharpe, L. (2000). Spectral sensitivities of the middle- and long-wavelength sensitive cones derived from measurements in observers of known genotype. *Vision Research*, *40*(13), 1711–1737.

Stockman, A., Sharpe, L. T., & Fach, C. C. (1999). The spectral sensitivity of the human short-wavelength cones. *Vision Research*, *39*(17), 2901–2927.

Thibos, L. N., Bradley, A., Still, D. L., Zhang, X., & Howarth, P. A. (1990). Theory and measurement of ocular chromatic aberration. *Vision Research*, *30*(1), 33–49.

Thibos, L. N., Hong, X., Bradley, A., & Cheng, X. (2002). Statistical variation of aberration structure and image quality in a normal population of healthy eyes. *Journal of the Optical Society of America A*, *19*(12), 2329–2348.

Tkacik, G., Garrigan, P., Ratliff, C., Milcinski, G., Klein, J. M., Sterling, P., … Balasubramanian, V. (2011). Natural images from the birthplace of the human eye. *PLoS ONE*, *6*(6:e20409).

Van Gelder, R. N., & Buhr, E. D. (2016). Ocular photoreception for circadian rhythm entrainment in mammals. *Annual Review of Vision Science*, *2*.

von Kries, J. (1902). Chromatic adaptation. In *Sources of color vision* (pp. 109–119). Cambridge, MA: MIT Press.

Vrhel, M., Gershon, R., & Iwan, L. (1994). Measurement and analysis of object reflectance spectra. *Color Research And Application*, *19*(1), 4–9.

Wald, I., Dietrich, A., Benthin, C., Efremov, A., Dahmen, T., Gunther, J., … Slusallek, P. (2006, September). Applying ray tracing for virtual reality and industrial design. In *2006 IEEE symposium on interactive ray tracing* (pp. 177–185). ieeexplore.ieee.org.

Wald, I., Purcell, T. J., Schmittler, J., Benthin, C. *et al*. (2003). Realtime ray tracing and its use for interactive global illumination. *Eurographics, State of the Art Reports*.

Wandell, B. A. (1995). *Foundations of vision*. Sunderland, MA: Sinauer.

Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, *17*(3), 10.

Watson, A., & Pelli, D. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception and Psychophysics*, *33*(2), 113–120.

Wei, W. (2018). Neural mechanisms of motion processing in the mammalian retina. *Annual Review of Vision Science*, *4*, 165–192.

Whitehead, A., Mares, J., & Danis, R. (2006). Macular pigment. A review of current knowledge. *Archives of Ophthalmology*, *124*(7), 1038–1045.

Wikipedia contributors. (2021, July). *Lytro.* https://en.wikipedia.org/w/index.php?title=Lytro&oldid=1032099081. (Accessed: July 8, 2021.)

Williams, D. R., MacLeod, D., & Hayhoe, M. (1981). Foveal tritanopia. *Vision Research*, *19*(9), 1341–1356.

Williams, D. R., Sekiguchi, N., Haake, W., Brainard, D. H., & Packer, O. (1991). The cost of trichromacy for spatial vision. In *From pigments to perception* (pp. 11–22). New York: Plenum Press.

Wyszecki, G. (1958). Evaluation of metameric colors. *Journal of the Optical Society of America*, *48*(7), 451–454.

Wyszecki, G., & Stiles, W. (1982). *Color science: Concepts and methods, quantitative data and formulae*, 2nd ed. New York: John Wiley & Sons.

Yellott, Jr., J. I., Wandell, B. A., & Cornsweet, T. N. (1984). The beginnings of visual perception: The retinal image and its initial encoding. In I. Darien-Smith (Ed.), *Handbook of physiology: The nervous system* (Vol. III, pp. 257–316). New York: Easton.

Young, T. (1802). On the theory of light and colours. *Philosophical Transactions of the Royal Society of London*, *92*, 20–71.

Zhang, L., Cottaris, N. P., & Brainard, D. H. (2021). An image reconstruction framework for characterizing early vision. *bioRxiv*. doi: 10.1101/2021.06.02.446829

2 Measuring Multisensory Integration in Selected Paradigms

Adele Diederich and Hans Colonius

2.1 Overview

The investigation of processes involved in merging information from different sensory modalities has become the subject of research in many areas, including anatomy, physiology, and behavioral sciences. This field of research termed "multisensory integration" (MI) is flourishing, crossing borders between psychology and neuroscience. The focus of this chapter is on *measures* of multisensory integration based on numerical data collected from single neurons and in behavioral paradigms: spike numbers, reaction time, frequency of correct or incorrect responses in detection, recognition, and discrimination tasks. Defining that somewhat fuzzy term, it has been observed that at least some kind of *numerical*

measurement assessing the strength of crossmodal effects is required. On the empirical side, these measures typically serve to quantify effects of various covariates on MI, like age, certain disorders (e.g., dyslexia), developmental conditions, training and rehabilitation, in addition to attention and learning. On the theoretical side, these measures often help to probe hypotheses about underlying integration mechanisms like optimality in combining information or inverse effectiveness, without necessarily subscribing to a specific model.

Given the important role of its neurophysiological basis, we start with a presentation of the major rules of integration observed in neural responses in the form of spike numbers elicited, and introduce numerical measures based on them. The essential role of the concept of "probability summation" in deriving measures satisfying certain "optimality" criteria emerges soon, and it reappears in later sections on measures based on response speed in different behavioral paradigms.[1]

Subsequently, measures based on accuracy are discussed in the context of signal detection theory, followed by measures developed within the broad area of audiovisual speech identification. A proposal for measuring integration efficiency based on the *Fechnerian scaling* approach closes that section.

The number of models trying to reveal the mechanisms underlying MI at different levels of description, from the neural to the behavioral, is large and growing. In the corresponding section, we had to be very selective, and we primarily sketch models that help motivate a specific measure of integration.

In order to keep the presentation focused, measures suggested for multisensory "illusions," like the McGurk effect or the sound-induced flash illusion (typically, percentages), are not considered at all, nor are those derived from functional magnetic resonance imaging data sets. A list of all measures discussed in the chapter is found in the discussion section. Finally, the reader should not expect a balanced presentation of the large field of measuring multisensory integration; instead, we mainly consider those more or less related to our own work.

2.2 Measures of Multisensory Integration: Introduction

2.2.1 Defining Multisensory Integration

Progress in MI is documented in several recent handbooks (see Section 2.8 for an overview of the literature). Due to the large range of contexts – from neurophysiology to applied psychology and marketing, from single cells to food tastes – the field has been labeled in different ways (e.g. as "intersensory facilitation/enhancement," "intersensory/crossmodal interaction," or "multisensory integration"), creating some semantic confusion among many researchers. In 2010, a group of authors, together with Barry Stein, one of the founders of the field in neuroscience, agreed upon defining "multisensory integration" as

1 Optimality is always defined here only in relation to a specific paradigm.

> the neural process by which unisensory signals are combined to form a new product. It is operationally defined as a multisensory response (neural or behavioral) that is significantly different from the responses evoked by the modality-specific component stimuli. (B. E. Stein *et al.*, 2010, p. 1719)

This broad definition does not commit to a specific model or experimental paradigm, nor to a criterion of optimality. Nevertheless, it requires some type of measure to assess whether the multisensory response is "significantly different" from the unisensory responses. Investigating such measures, as well as some models related to them, is the focus of this chapter.[2] Moreover, while the definition encompasses both facilitation and inhibition of the multisensory response, most measures presented here are formulated for the case of facilitation only and would need to be adapted to comprise inhibition.

2.2.2 Measuring Multisensory Integration

First, we introduce some needed notation (see Table 2.1 for a list of abbreviations used in this chapter). Beginning with the stimulus side, stimuli of a specific modality are labeled by s_A, s_V, s_T, for auditory, visual, and tactile (or somatosensory)

Table 2.1 *Abbreviations used in the chapter.*

Acronym	Meaning
CRE	crossmodal response enhancement
E	expected value (mean)
FS	Fechnerian scaling
FLMP	fuzzy logical model of perception
IE	integration efficiency
MI	multisensory integration
OUP	Ornstein–Uhlenbeck process
PRE	prelabeling (model)
PS	probability summation
RMI	race model inequality
RT	reaction time
SC	superior colliculus
SDT	signal detection theory
SFE	statistical facilitation effect
SOA	stimulus onset asynchrony
SRT	saccadic reaction time
TOJ	temporal order judgment
TWIN	time window of integration (model)
UI	unisensory balance
VE/AE	visual/auditory enhancement

2 Note, however, that issues of testing *statistical* significance are not central to this chapter.

stimuli, respectively, where further stimulus-specific information, like intensity, has to be added as needed. When a label is only used as index of modality, we often omit the s part. A basic distinction to keep in mind is between a unisensory context where stimuli of a single modality, s_A, s_V, s_T, are presented, and a cross-sensory context where stimuli from two or more modalities are presented in a to-be-specified spatio-temporal arrangement. For concreteness, we refer to A, V, T as the unisensory context where only auditory, visual, or tactile stimuli are presented, respectively. Similarly, *VA* denotes a bisensory (visual–auditory) context with stimulus combinations labeled s_{VA} being presented, *VAT* a trisensory context with combined stimuli s_{VAT}, etc., where again further information about the specific presentation mode may have to be added. When the number of sensory modalities is not specified, we also use the label *crossmodal* (for context, condition, stimulus, response, etc.). Moreover, in this chapter mainly measures combining the visual and auditory modalities will be considered, but most of these would also apply to other modality combinations with minor modification.

Each time a specific auditory stimulus s_A, say, is presented, it will give rise to a unisensory response (e.g., a reaction time or a number of spikes within a certain time interval). Typically, these responses are considered as instantiation (realization) of some random variable (e.g., RT_A or N_A, respectively). Similarly, a combination stimulus s_{VA} elicits bisensory responses considered as realizations of some random variables, RT_{VA} or N_{VA}. To simplify the exposition, we will neglect all experimental details for now.

At the sample level, a descriptive measure of MI has to relate the set of multisensory responses to the sets of unisensory responses; for example, how much does the average auditory–visual response differ from the average auditory and average visual response? At the level of random variables, the MI measure should assess how, or how much, the distribution of responses to bisensory stimuli differs from the distributions to unisensory stimuli.

We define measures only at the level of probability distributions, the corresponding sample level measures are then easily derivable. In order to reduce the number of possible formats, one should consider necessary or desirable features of such a measure, denoted by CRE (crossmodal response enhancement/inhibition). We first state a few elementary properties any CRE measure of MI should have. The following list seems uncontroversial:

(i) (*Real-valued function*) CRE is a real-valued function of the crossmodal and unisensory empirical distributions, or of some parameter of these distributions (e.g., the mean).
(ii) (*No-integration case*) If the crossmodal distribution does not differ from one of the unisensory distributions, CRE equals zero.
(iii) (*Facilitation-inhibition*) Negative values of CRE indicate crossmodal inhibition, positive values crossmodal facilitation.

Clearly, these features do not impose strong restrictions on the form of the measure; this does not come as a surprise, however, given the huge number of

different experimental paradigms where MI is observable in various forms. Thus, (i) to (iii) should be seen as a minimal set of necessary requirements. Next, we consider two first examples satisfying them.

Example 2.1 (Spike numbers) The following measure of MI in a single neuron is common in neurophysiology:

$$\mathrm{CRE}_{SP} = \frac{\mathrm{E}N_{VA} - \max\{\mathrm{E}N_V, \mathrm{E}N_A\}}{\max\{\mathrm{E}N_V, \mathrm{E}N_A\}} \times 100, \tag{2.1}$$

where $\mathrm{E}N_{VA}$ is the mean[3] (absolute) number of spikes in response to the crossmodal stimulus and $\mathrm{E}N_V, \mathrm{E}N_A$ denote the mean (absolute) numbers of spikes to the visual and auditory unisensory stimuli, respectively.[4] Thus, CRE_{SP} quantifies crossmodal enhancement/inhibition as the percentage difference between the response to a crossmodal pair *VA* and the largest response to one of its unisensory components, *V* or *A*.

Example 2.2 (Reaction time measure) An analogous measure for RTs is

$$\mathrm{CRE}_{RT} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\} - \mathrm{E}RT_{VA}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\}} \times 100, \tag{2.2}$$

where $\mathrm{E}RT_{VA}$ is mean RT to an auditory–visual stimulus combination and $\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\}$ is the faster of the unisensory mean RTs to the visual and auditory stimulus. Thus, CRE_{RT} expresses multisensory enhancement/inhibition as a proportion of the faster unisensory response. For example, $\mathrm{CRE}_{RT} = 10$ means that mean response time to the visual–auditory stimulus is 10% faster than the faster of the expected response times to unimodal visual and auditory stimuli.

2.3 Measures for the Multisensory Neuron Response

2.3.1 Rules of Multisensory Integration

The first systematic neuronal studies of MI, performed in the 1970s, focused on a midbrain structure, the cat *superior colliculus* (SC) (Meredith & Stein, 1983). Stein and colleagues showed that neurons in the deep layers of the SC are primary sites of multisensory convergence: if a visual–auditory stimulus combination is presented such that the visual stimulus is within its visual receptive field and the auditory stimulus is within its auditory receptive field, it will typically produce response enhancement, in the form of increased spike numbers, even when the stimuli are not found at the exact same spatial location. Likewise, response depression (inhibition) tends to occur if the visual stimulus is within its receptive field while the auditory stimulus is outside its receptive field. This has become known as the *spatial rule* of MI.

3 Note that we drop the brackets in E[.] when there is no risk of confusion.

4 Spike numbers are counted in a specified time interval and may or may not include spontaneous activity.

Similarly, changing the interval between auditory and visual stimulation can change enhancement to depression: presenting a visual stimulus 50 ms or 150 ms before the auditory (*V*50*A* or *V*150*A*, for short) produced response enhancement, whereas longer intervals (*V*300*A* or *A*200*V*) produced fewer impulses than a unisensory stimulus (i.e., depression) (Meredith & Stein, 1983). The effect, termed *temporal rule* of MI, largely depends on the amount of overlap of the peak discharge periods of the neuron's unisensory responses. Later, these spatiotemporal rules of single neuron recordings have also been observed in other species like the monkey, ferret, owl, guinea pig, rat, snake, and others.

A third major factor affecting MI is the efficacy of the component stimuli within the neuronal receptive fields. Response enhancement is found to be greater the less effective the unisensory stimuli are. This rule of *inverse effectiveness* is most impressive when the unisensory stimulus intensities are below the threshold of eliciting any response from the neuron but in combination generate a reliable response.

More recently, a more nuanced function of unisensory signal strength and the temporal rule has been observed in cat SC (R. Miller *et al.*, 2015). For each neuron, response magnitude (mean number of impulses per trial) to the visual (*V*) and the auditory stimuli (*A*) can be used to quantify the notion of *unisensory imbalance* (UI):

$$\mathrm{UI} = \frac{|\mathrm{E}N_A - \mathrm{E}N_V|}{\mathrm{E}N_A + \mathrm{E}N_V} \times 100. \tag{2.3}$$

UI quantifies the relative difference between the response magnitude to the visual and the auditory stimuli. It has a minimum of zero when the visual and auditory responses are of equal magnitude and a maximum of 100 when one of the responses is lacking.

In view of the above definition of crossmodal enhancement [Equation (2.1)], increasing unisensory imbalance should not affect CRE_{SP}. However, across a wide range of response magnitude, increasing imbalance was found to be coupled with both a decrease in the multisensory response ($\mathrm{E}N_{VA}$) and in crossmodal enhancement CRE_{SP} (see Figure 2.1). Moreover, the order of arrival also mattered: when the unisensory response magnitudes were imbalanced, multisensory enhancement was maximized when stronger responses were advanced in time relative to weaker responses ("stronger first") and minimized when stronger responses were delayed ("stronger second") (for details, see R. Miller *et al.*, 2015). Thus, only when the unisensory stimuli are "balanced" does multisensory enhancement depend solely on their absolute temporal offset.

Still a different twist on the single-cell mechanism in SC has emerged from developmental findings. Since the early studies, it had been known that, just before and after birth, cat SC neurons are largely unresponsive to sensory stimulation and lack spontaneous activity. Successively, neurons start responding to tactile, then auditory, and finally visual stimulation. Besides unisensory neurons, multisensory neurons appear, but they do not yet show enhanced responses, instead they appear

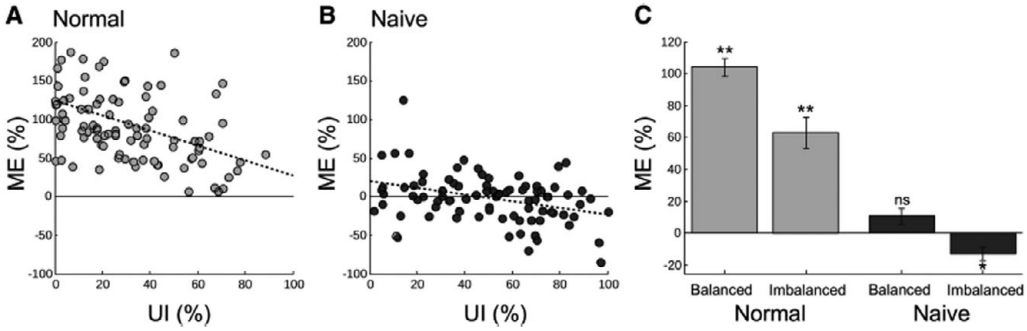


Figure 2.1 *Relationships between multisensory responses* (ME ≡ CRE_{SP}) *and unisensory imbalance (UI) in normal and naïve cohorts. (A) Neurons from normally reared animals produce their greatest response enhancements when the spatiotemporally concordant cues produced balanced unisensory responses: an inverse relationship between ME and UI (dotted line). (B) Naïve SC neurons showed a similar inverse relationship between ME and UI, but even-balanced samples failed to produce significantly enhanced multisensory products, and imbalanced samples induced multisensory depression. (C) Histograms summarizing the results. Vertical lines through the bars represent standard error (from Yu* et al.*, 2019).*

to act as a common conduit for different senses to reach the same motor output systems. These early studies had shown that blocking an animal's multisensory experience (e.g., rearing cats with no visual stimulation at all) results in multisensory responses not stronger than the most effective component, suggesting CRE to be equal to zero. However, findings by Yu and colleagues (Yu *et al.*, 2019) revealed that there exists competition between the senses in these "naïve" neurons: crossmodal stimuli, whether spatio-temporally disparate or not, can elicit inhibition in these neurons' responses. They conclude that the default mode of multisensory processing in SC is competition rather than absence of integration, and they develop a neurocomputational model consistent with this assumption. Thus, some form of MI (including competition) seems to occur at all stages of maturation, and the ability of enhanced (orienting) responses to crossmodal events increases over subsequent stages of development (Yu *et al.*, 2019, p. 1374).

All the rules, sometimes referred to as *principles* of MI, discussed above have raised a discussion about whether, and in how far, they also determine multisensory behavior in humans and under more complex stimulus contexts. Before we follow up on these issues, we need to consider an aspect that proved particularly noteworthy in measuring MI.

2.3.2 Multisensory Integration vs. Probability Summation

The fact that a multisensory neuron is responsive to multiple sensory modalities does not guarantee that it has actually engaged in integrating its multiple sensory

inputs. Rather, it may simply respond to the most effective stimulus in a given trial (i.e., to the stimulus eliciting the strongest response).[5] In other words, it is possible that the response to a visual–auditory stimulus is simply determined by the larger of the responses to the modality-specific components, that is, by the component that happens to elicit the higher absolute number of spikes in a given trial. Assuming random variation of the responses, such a mechanism is known as *probability summation* (PS).

In order to explore implications for how to measure MI in single neurons in the presence of PS, we first introduce some relevant statistical concepts. Only the case of facilitation will be discussed here, while the case of inhibition can be developed analogously. As before, the unisensory (visual, auditory) responses are conceived of as realizations of random variables N_V and N_A. We define distribution functions G_V and G_A, respectively:

$$\mathrm{P}[N_V \leq n_V] = G_V(n_V) \quad \text{and} \quad \mathrm{P}[N_A \leq n_A] = G_A(n_A),$$

with n_V and n_A taking integer values $0, 1, \ldots$. For the bisensory condition, we assume a distribution function G_{VA} exists such that

$$\mathrm{P}[N_{VA} \leq n] = G_{VA}(n),$$

with $n = 0, 1, \ldots$. Thus, N_V, N_A, and N_{VA} are random variables whose realizations (samples) are observed in the experiment under to-be-specified conditions.

Probability Summation (PS) in Spike Numbers

For clarity, the three assumptions underlying the concept of PS in this multisensory context will be stated in detail. The first assumption refers to the observation that realizations of the random variables N_V and N_A are collected under different stimulus conditions (visual vs. auditory) and, thus, occur in distinct probability spaces. A priori, there is no prescribed way to combine them. In particular, any assumption about stochastic (in-)dependence between N_V and N_A is meaningless. However, one can postulate a *stochastic coupling*[6] of the two random variables.

Assumption 1: There exists a random vector $(\tilde{N}_V, \tilde{N}_A)$ *with a joint distribution* $\tilde{H}_{VA}$*:*

$$\tilde{H}_{VA}(n_V, n_A) = \mathrm{P}[\tilde{N}_V \leq n_V, \tilde{N}_A \leq n_A].$$

Assuming the existence of $\tilde{H}_{VA}$ amounts to a *coupling* of the random variables $\tilde{N}_V$ and $\tilde{N}_A$, which is always possible. Of course, we want $\tilde{N}_V$ and $\tilde{N}_A$ to be a "copy" of N_V and N_A in the following sense.

5 As Stein and colleagues (B. E. Stein *et al.*, 2009, p. 114) have put it, "At the time of the early physiology studies in the 1980s, it was considered possible that these neurons only represented a common route by which independent inputs from a variety of senses could gain access to the same motor apparatus in generating behavior (e.g., possibly employing a 'winner-take-all' algorithm)."

6 See Colonius (2016) for an introduction to the concept in this context.

Assumption 2: The marginal distributions of $\tilde{H}_{VA}(n_V, n_A)$ are equal to G_V and G_A, respectively:

$$\tilde{H}_{VA}(n_V, \infty) = G_V(n_V) \quad \textit{and} \quad \tilde{H}_{VA}(\infty, n_A) = G_A(n_A).$$

This important restriction, equating the marginals to the observable unisensory response distributions, is often called "context invariance."

Note that we have not assumed a specific form for $\tilde{H}_{VA}$. In fact, we are only interested in the values on the diagonal, $\tilde{H}_{VA}(n,n)$. For $n = 0, 1, \ldots$, we write

$$\begin{aligned}\tilde{H}_{VA}(n,n) &= \mathrm{P}\left[\{\tilde{N}_V \leq n\} \cap \{\tilde{N}_A \leq n\}\right] \\ &= \mathrm{P}\left[\max\{\tilde{N}_V, \tilde{N}_A\} \leq n\right] \\ &\equiv \tilde{G}_{VA}(n).\end{aligned}$$

The third assumption specifies the probability mechanism proper.

Assumption 3: For $n = 0, 1, \ldots$

$$G_{VA}(n) = \tilde{G}_{VA}(n). \tag{2.4}$$

That is, the observable crossmodal responses are the result of taking the maximum of the unisensory responses.

It is always possible to construct *some* bivariate distribution $\tilde{H}_{VA}(n_V, n_A)$ (e.g., by assuming stochastic independence):

$$\tilde{H}_{VA}(n_V, n_A) = \mathrm{P}\left[\tilde{N}_V \leq n_V\right] \mathrm{P}\left[\tilde{N}_A \leq n_A\right],$$

which implies the empirically testable hypothesis

$$G_{VA}(n) = \tilde{G}_{VA}(n) = G_V(n)\, G_A(n)$$

for $n = 0, 1, \ldots,$ under context invariance (*Assumption 2*).

In general, however, it is not obvious how *Assumption 3* should be tested. Stochastic independence, while convenient, may not be the most judicious choice, as will be argued below.

2.3.3 Measures of MI under PS Hypothesis

It is straightforward to compare observed responses with those predicted by PS: one has to gauge the difference between the means (expected values) associated with G_{VA} and $\tilde{G}_{VA}$, that is $\mathrm{E}N_{VA}$ and $\mathrm{E}\max\{N_V, N_A\}$, respectively. The common measure of MI based on spike counts introduced in Example 2.1:

$$\mathrm{CRE}_{SP} = \frac{\mathrm{E}N_{VA} - \max\{\mathrm{E}N_V, \mathrm{E}N_A\}}{\max\{\mathrm{E}N_V, \mathrm{E}N_A\}} \times 100 \tag{2.5}$$

is then replaced by

$$\mathrm{CRE}^*_{SP} = \frac{\mathrm{E}N_{VA} - \mathrm{E}\max\{N_V, N_A\}}{\mathrm{E}\max\{N_V, N_A\}} \times 100. \tag{2.6}$$

Note that *Assumption 2* permits us to write measure CRE^*_{SP} with N_V, N_A instead of $\tilde{N}_V, \tilde{N}_A$. By a well-known statistics result (*Jensen's inequality*; e.g., Ross, 1996):

$$\max\{\mathrm{E}N_V, \mathrm{E}N_A\} \leq \mathrm{E}\max\{N_V, N_A\}$$

always holds, obviously implying

$$\mathrm{CRE}^*_{SP} \leq \mathrm{CRE}_{SP}. \tag{2.7}$$

This inequality reveals an important consequence: in order to assess "true" MI, that is, over and above the effect of PS, the criterion *mean* number of spikes observed ($\mathrm{E}N_{AV}$) has to be larger than the mean taking PS into account.

Effects of Unisensory Imbalance

The move from CRE_{SP} to CRE^*_{SP} opens up the possibility to probe effects of unisensory imbalance mentioned above [Equation (2.3)]:

$$\mathrm{UI} = \frac{|\mathrm{E}N_A - \mathrm{E}N_V|}{\mathrm{E}N_A + \mathrm{E}N_V}.$$

Note that only the maximum of $\mathrm{E}N_A$ and $\mathrm{E}N_V$ enters into CRE_{SP}, so that varying imbalance has no effect on that index. In contrast, computing $\mathrm{E}\max\{N_V, N_A\}$ involves the distribution of both variables, N_A and N_V, and it is easy to find instances where CRE^*_{SP} depends on both $\mathrm{E}N_A$ and $\mathrm{E}N_V$ simultaneously (see, e.g., Colonius & Diederich, 2017 for an example with Poisson-distributed spike counts).

Towards an Optimal Measure of MI

Inequality (2.7) holds without assuming a specific distribution for $\tilde{G}_{VA}$. While stochastic independence between N_V and N_A is typically taken for granted in computing the value of $\mathrm{E}\max\{N_V, N_A\}$, it turns out that it is not the most conservative choice possible.[7] To demonstrate, we recall (without proof) a classic result from statistics (Fréchet, 1951) about upper and lower bounds for arbitrary distributions, here applied to $\tilde{H}_{VA}$.

Lemma 2.3 (Fréchet inequalities) *For $m, n = 0, 1, \ldots$, let $\tilde{H}_{VA}(m,n) = P(\tilde{N}_V \leq m, \tilde{N}_A \leq n)$ be a bivariate distribution with marginals $\tilde{G}_V(m), \tilde{G}_A(n)$, respectively. Then*

$$\max\{0, \tilde{G}_V(m) + \tilde{G}_A(n) - 1\} \leq \tilde{H}_{VA}(m,n) \leq \min\{\tilde{G}_V(m), \tilde{G}_A(n)\}.$$

The upper and lower bound in the lemma represent bivariate distributions as well, with the same marginals as $\tilde{H}_{VA}(m,n)$ but possessing maximal positive, respectively negative, dependence between $\tilde{N}_V$ and $\tilde{N}_A$ (e.g., Joe, 1997). Setting $m = n$, we denote the lower bound with maximal negative dependence by $\tilde{G}^{(-)}_{VA}(n)$. Then,

$$\tilde{G}^{(-)}_{VA}(n) \equiv \max\{0, \tilde{G}_V(n) + \tilde{G}_A(n) - 1\} \leq \tilde{G}_{VA}(n) \tag{2.8}$$

for $n = 0, 1, \ldots$.

7 Here, "conservative" means that one wants to avoid claiming MI to hold when, in reality, it does not.

Importantly, maximal negative dependence between $\tilde{N}_V$ and $\tilde{N}_A$ maximizes the expected value of $\mathrm{E}\max\{N_V, N_A\}$:

Lemma 2.4. *Let* $E^{(-)}\max\{\tilde{N}_V, \tilde{N}_A\}$ *be the expected value of* $\max\{\tilde{N}_V, \tilde{N}_A\}$ *under bivariate distribution* $\max\{0, \tilde{G}_V(m) + \tilde{G}_A(n) - 1\}$*; then*

$$E\max\{\tilde{N}_V, \tilde{N}_A\} \leq E^{(-)}\max\{\tilde{N}_V, \tilde{N}_A\}$$

under any bivariate distribution $\tilde{H}_{VA}(m,n)$ *for* $E\max\{\tilde{N}_V, \tilde{N}_A\}$.

This can be shown as follows. Rewriting Equation (2.8) as

$$1 - \tilde{G}_{VA}(n) \leq 1 - \tilde{G}_{VA}^{(-)}(n)$$

and summing over all n yields

$$\mathrm{E}\max\{N_V, N_A\} \equiv \sum_{n=0}^{\infty}[1 - \tilde{G}_{VA}(n)] \leq \sum_{n=0}^{\infty}[1 - \tilde{G}_{VA}^{(-)}(n)] \equiv \mathrm{E}^{(-)}\max\{N_V, N_A\}.$$

The upshot of Lemma 2.4 is that an optimal choice for defining CRE_{SP}^{*} [Equation (2.6)] is to insert $\mathrm{E}^{(-)}\max\{N_V, N_A\}$:

Definition 2.5. The measure of MI taking into account PS with maximal negative dependence between the unisensory responses is

$$\mathrm{CRE}_{SP}^{max} = \frac{\mathrm{E}N_{VA} - \mathrm{E}^{(-)}\max\{N_V, N_A\}}{\mathrm{E}^{(-)}\max\{N_V, N_A\}} \times 100. \tag{2.9}$$

Note that it is not claimed here that a multisensory neuron actually operates under this extreme negative dependency rule. As long as PS is considered a possible alternative to "true" MI, however, some specification of the stochastic relation between the unisensory responses has to be made in CRE_{SP}^{*}. Assuming maximal negative dependency is simply the most efficient way to hedge against a "false alarm," that is, declaring true MI while enhancement may simply be a product of PS. Whenever there is empirical or theoretical evidence in favor of some other form of dependence (e.g., stochastic independence), this could be used to modify the benchmark appropriately.

Because, in general, the new measure is more restrictive than the traditional CRE measure, many neurons previously categorized as "multisensory" may lose that property. The purpose of the new measure corresponds to that of the traditional measure: given a fixed statistical criterion, one may categorize a single neuron as either being "multisensory" or not. It is, of course, possible that a neuron actually "truly" integrates the unimodal activations but still does not meet the criterion set by maximal negative PS. However, as long as one has no direct insight into the integration mechanism, an alternative interpretation in terms of PS simply cannot be ruled out.

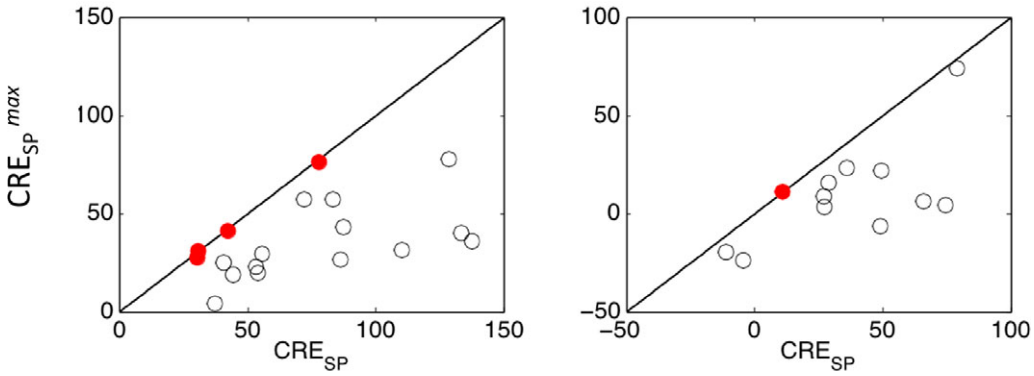


Figure 2.2 *Pairs of sample estimates of* $(\mathrm{CRE}_{SP}, \mathrm{CRE}_{SP}^{max})$ *based on 27 recording blocks (15 stimulus presentations in each block). In the left-hand panel spontaneous activity was included, in the right-hand panel it has been removed. Filled circles indicate no significant difference between* CRE_{SP} *and* CRE_{SP}^{max}*, based on bootstrap confidence intervals (*$N = 10{,}000$, $\alpha = 0.05$*). Thus, each open circle refers to a recording where the label multisensory may be lost when applying measure* CRE_{SP}^{max}*. There were 4 out of 27 cases with no significant difference between both measures (left panel), after spontaneous activity was removed, only 1 out of 19 cases was not significant (right panel) (from Colonius & Diederich, 2017).*

Example Application of CRE_{SP}^{max}

Estimating $\mathrm{E}\max\{\tilde{N}_V, \tilde{N}_A\}$ from sample data is straightforward. Without going into detail, the procedure is as follows. We have two samples of numbers of spikes from each modality of size n_v and n_a, say, and assume $n_v = n_a$. Under the stochastic independence version of SP, the number of spikes occurring in trial $i, i = 1, \ldots, n_v$ is randomly paired with the number of spikes in trial $j, j = 1, \ldots, n_a$ (without replacement). The maximum in each pair is determined and the average of the maxima yields an estimate of $\mathrm{E}\max\{\tilde{N}_V, \tilde{N}_A\}$.

Under maximal negative dependence of PS, trial i with the largest number of spikes is paired with trial j with the smallest number of spikes, the second largest i is paired with the second lowest j, and so on (method of "antithetic variables"), and the average of the maxima is again computed as the estimate of $\mathrm{E}\max\{\tilde{N}_V, \tilde{N}_A\}$. If the unisensory samples are of different sizes, some replacement procedure could be applied. In an illustrative sample of cat SC neurons,[8] Colonius and Diederich (2017) showed that there was a significant decrease from CRE_{SP} to CRE_{SP}^{max} in 24 out of 27 recording blocks collected from 20 neurons. Whether or not the label "multisensory" is actually lost for some neurons, however, depends on criteria of the statistical test comparing the sample means (see Figure 2.2).

8 Data provided by the lab of Mark Wallace (personal communication).

2.4 Measures Based on Response Speed

The earliest observations of MI effects have likely been reported in the context of measuring the speed and accuracy of responses to crossmodal stimuli at the beginning of the twentieth century (see Welch & Warren, 1986, for a review). In a typical paradigm, participants are instructed to respond via button press as soon as a signal of any modality occurs (*redundant signals paradigm*[9]). It is to be distinguished from a related paradigm, often called *focused attention paradigm*; in the latter, one modality is designated as "target" modality, the other as "distractor" modality, and participants are instructed to respond only to signals from the target modality (mostly, visual) but not to distractor signals. The two paradigms demand separate treatments for the measurement of MI.

Note that erroneous responses should also be defined differently for the two paradigms, but we will first ignore errors entirely since they are often kept at a negligible rate in the experiments. Accuracy measures are discussed later.

2.4.1 MI Measures in Redundant Signals Paradigms

In general, bisensory, in particular visual–auditory, stimulation results in smaller mean RT compared to unisensory stimulation, and responses to trisensory stimulation (often visual, auditory, and tactile) are faster on average than to bisensory stimulation. The magnitude of the speed-up depends on the specifics of the experiment, in particular the intensity of the different modalities and their temporal configuration. For visual–auditory presentations, the greatest effect is typically found when the visual stimulus precedes the auditory by an interval that equals the difference between the unisensory mean RTs.

Hence the MI measure for RTs introduced in Equation (2.2) should be augmented to include stimulus onset asynchrony (SOA), denoted as τ:

$$\mathrm{CRE}_{RT,\tau} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau\} - \mathrm{E}RT_{V\tau A}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau\}} \times 100, \tag{2.10}$$

where $RT_{V\tau A}$ is the RT to a visual–auditory stimulus combination with the visual preceding the auditory by τ [ms]; thus, the maximum of $\mathrm{CRE}_{RT,\tau}$ would be expected[10] for $\tau = \mathrm{E}RT_V - \mathrm{E}RT_A$.

For trisensory stimulus contexts (*VAT*), the analogous measure is

$$\mathrm{CRE}_{RT,\tau_1\tau_2} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau_1, \mathrm{E}RT_T + \tau_1 + \tau_2\} - \mathrm{E}RT_{V\tau_1 A\tau_2 T}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau_1, \mathrm{E}RT_T + \tau_1 + \tau_2\}} \times 100, \tag{2.11}$$

where $RT_{V\tau_1 A\tau_2 T}$ is the RT to a visual–auditory–tactile stimulus combination with the visual preceding the auditory by τ_1 [ms] and the auditory preceding the tactile by τ_2 [ms].

9 Also known as *divided attention paradigm*.

10 Visual RTs tend to be slower than auditory RTs at comparable intensity levels.

Note that adding a third modality increases the possible measures of response enhancement: trisensory response speed may now also be compared with the speed of any bisensory combination (e.g., $V\tau_1 A\tau_2 T$ with $V\tau_1 A$ or $A\tau_2 T$), as long as these different combinations have been presented in the experiment. For example:

$$\mathrm{CRE}_{RT,(\tau_1)\tau_2} = \frac{\mathrm{E}RT_{V\tau_1 A} - \mathrm{E}RT_{V\tau_1 A\tau_2 T}}{\mathrm{E}RT_{V\tau_1 A}} \times 100, \tag{2.12}$$

measuring the additional multisensory effect of a tactile stimulus, presented τ_2 [ms] later, on the speed of a visual–auditory combination.

2.4.2 Probability Summation in the Redundant Signals Paradigm

None of the RT measures of MI considered so far takes the PS hypothesis into account. In this context, the hypothesis amounts to postulating the so-called *race model* and as such, arguably, represents the most widely known version of PS in multisensory research. The idea is that, for example, a visual–auditory stimulus combination triggers random visual and auditory processing times such that the observed RT equals the minimum of the two (i.e., the "winner of the race").

Usually, RTs are assumed to comprise some additive components, like motor preparation and execution. To simplify the discussion, we neglect this distinction here. Observed samples from random variables, denoted as T_V, T_A, and T_{VA}, represent RTs obtained in unisensory visual, auditory, and bisensory trials, respectively. Thus, we equate realizations of T_V, T_A, and T_{VA} with the observable RT under these conditions.

We define underlying distribution functions F_V and F_A, respectively:

$$\mathrm{P}[T_V \leq t_V] = F_V(t_V) \quad \text{and} \quad \mathrm{P}[T_A \leq t_A] = F_A(t_A),$$

with T_V and T_A taking on non-negative real numbers. For the bisensory context, we assume a distribution function F_{VA} such that

$$\mathrm{P}[T_{VA} \leq t] = F_{VA}(t),$$

with $t \geq 0$. Hence, T_V, T_A, and T_{VA} are random variables whose realizations are observed in an experiment under to-be-specified conditions.

Probability Summation (PS) in Reaction Times

The exact definition of PS follows in close analogy to the one given for spike numbers in the previous section.

Assumption 1: There exists a random vector $(\tilde{T}_V, \tilde{T}_A)$ with a joint distribution $\tilde{K}_{VA}$:

$$\tilde{K}_{VA}(t_V, t_A) = \mathrm{P}[\tilde{T}_V \leq t_V, \tilde{T}_A \leq t_A].$$

Assuming the existence of $\tilde{K}_{VA}$ amounts again to a *coupling* of the random variables $\tilde{T}_V$ and $\tilde{T}_A$, which is always possible. Of course, we want $\tilde{T}_V$ and $\tilde{T}_A$ to be a "copy" of T_V and T_A in the following sense.

Assumption 2: *The marginal distributions of $\tilde{K}_{VA}(t_V, t_A)$ are equal to F_V and F_A, respectively:*

$$\tilde{K}_{VA}(t_V, \infty) = F_V(t_V) \quad \textit{and} \quad \tilde{K}_{VA}(\infty, t_A) = F_A(t_A).$$

Thus, "context invariance" is postulated for RT distributions as well. It follows that for $t \geq 0$:

$$\begin{aligned}\tilde{K}_{VA}(t,t) &= \mathrm{P}\left[\{\tilde{T}_V \leq t\} \cap \{\tilde{T}_A \leq t\}\right] \\ &= \mathrm{P}\left[\max\{\tilde{T}_V, \tilde{T}_A\} \leq t\right] \\ &\equiv \tilde{F}_{VA}(t).\end{aligned}$$

Assumption 3: *For $t \geq 0$*

$$F_{VA}(t) = \tilde{F}_V(t) + \tilde{F}_A(t) - \tilde{F}_{VA}(t). \tag{2.13}$$

Assumption 2 is the central one again, implying that the observable crossmodal RTs result from taking the minimum of the unisensory RTs (race model).

It is always possible to construct some bivariate distribution $\tilde{K}_{VA}(t_V, t_A)$ (e.g., by assuming stochastic independence):

$$\begin{aligned}\tilde{K}_{VA}(t_V, t_A) &= \mathrm{P}\left[\tilde{T}_V \leq t_V\right] \mathrm{P}\left[\tilde{T}_A \leq t_A\right] \\ &= F_V(t_V)\, F_A(t_A) \qquad \text{by } \textit{Assumption 2}.\end{aligned}$$

This implies the special case of "independent race model":

$$F_{VA}(t) = 1 - (1 - F_V(t))(1 - F_A(t)). \tag{2.14}$$

The PS hypothesis has been studied as a possible non-parametric model for RTs in the redundant signals paradigm. Being equivalent to the "race model," it predicts a specific relation between the distribution functions for bisensory and unisensory conditions:

$$\begin{aligned}F_{VA}(t) &= \tilde{F}_V(t) + \tilde{F}_A(t) - \tilde{F}_{VA}(t) \qquad \text{by } \textit{Assumption 3} \\ &= F_V(t) + F_A(t) - \tilde{F}_{VA}(t) \qquad \text{by } \textit{Assumption 2} \\ &\leq F_V(t) + F_A(t). \end{aligned} \tag{2.15}$$

Inequality (2.15) is a simple version of *Boole's inequality* and has been called "race-model inequality" (RMI) in this context. Testing it has become routine in a vast number of empirical studies, using a variety of different statistical procedures.[11] Note that the right-hand side of RMI approaches 2 for t going to infinity, so it can be replaced by $\min\{F_V(t) + F_A(t), 1\}$. Typically, RMI tends to be violated for t not too large.

11 Sometimes, the "independent" version of the inequality is tested, $F_{VA}(t) \leq F_V(t) + F_A(t) - F_V(t)F_A(t)$, but violations of this inequality would only rule out the special case of a stochastically independent race.

2.4.3 Measures of MI in Redundant Signals Paradigms under PS

In addition to testing the race model, a quantitative measure of the degree of RMI violation has been proposed. The latter turns out to be the basis of a measure of MI in redundant signal experiments.

We define a function $R_{VA}(t)$, for $t \geq 0$:

$$R_{VA}(t) \equiv F_{VA}(t) - \min\{F_V(t) + F_A(t), 1\}. \tag{2.16}$$

Hence, values of t with $R_{VA}(t) > 0$ indicate a violation of RMI, whereas values of t with $R_{VA}(t) \leq 0$ are compatible with the race model. The positive part of the area between $F_{VA}(t)$ and $\min\{F_V(t) + F_A(t), 1\}$ is often taken as a measure of the amount of RMI violation. Integrating $R_{VA}(t)$ results in a convenient interpretation as MI measure. First, observe that

$$\begin{aligned} R_{VA}(t) &= F_{VA}(t) - \min\{F_V(t) + F_A(t), 1\} \\ &= 1 - \min\{F_V(t) + F_A(t), 1\} - [1 - F_{VA}(t)] \\ &= \max\{1 - F_V(t) - F_A(t), 0\} - [1 - F_{VA}(t)]. \end{aligned}$$

Integrating yields

$$\begin{aligned} \int_0^\infty R_{VA}(t)\, dt &= \int_0^\infty \max\{1 - F_V(t) - F_A(t), 0\}\, dt - \int_0^\infty [1 - F_{VA}(t)]\, dt \\ &= \mathrm{E}^{(-)} \min\{T_V, T_A\} - \mathrm{E}\{T_{VA}\}, \end{aligned}$$

where $\mathrm{E}^{(-)} \min\{T_V, T_A\}$ denotes mean RT predicted by a race model with maximal negative dependence between the latencies T_V and T_A. This leads to a modified version of $\mathrm{CRE}_{RT,\tau}$ [see Equation (2.10)] accounting for PS:

$$\mathrm{CRE}^{min}_{RT,\tau} = \frac{\mathrm{E}^{(-)} \min\{RT_V, RT_A + \tau\} - \mathrm{E}RT_{V\tau A}}{\mathrm{E}^{(-)} \min\{RT_V, RT_A + \tau\}} \times 100, \tag{2.17}$$

where T_V, T_A are identified with RT_V, RT_A, respectively.

2.4.4 MI Measures in Focused Attention Paradigms

Let us assume a stimulus from the visual modality is the target. The task is to respond to the occurrence of the target, via button press, while ignoring an auditory stimulus ("distractor") presented in spatio-temporal proximity. In a frequent variant, the required response is to execute an eye movement towards a target that occurs at a randomized spatial position in the visual field, with saccadic RT and/or accuracy of the trajectory/landing position being recorded. In all cases, MI is measured by how much the response to the target is modulated by the presence of a distractor. For RTs, a simple adaption of the CRE measure in the redundant paradigm results in

$$\mathrm{CRE}_{RT} = \frac{\mathrm{E}RT_V - \mathrm{E}RT_{VA}}{\mathrm{E}RT_V} \times 100. \tag{2.18}$$

The amount and direction (facilitation vs. inhibition) of CRE_{RT} depends on a host of experimental conditions. Because visual and auditory stimuli activate visuomotor neurons in superior colliculus (SC), thereby eliciting goal-directed eye movements, many studies of MI have focused on gaze behavior, in particular saccadic reaction time.[12]

While the temporal and spatial rules of MI are, in general, consistent with findings in the redundant signals task, effects of the role of localizability of the auditory distractor have found special attention in eye movement experiments. Specifically, when target and distractor are presented at the same position (e.g., both above or below fixation point), SRTs are faster than when they are presented at opposite positions (e.g., target above, distractor below fixation point). However, this effect disappears when localization of the auditory stimulus is made more difficult (e.g., by increasing the level of a background noise). Hence, the *perceived* rather than the physical distance between target and distractor controls the MI effect (Colonius, Diederich, & Steenken, 2009).

2.5 MI Measures Based on Accuracy

Next, we discuss MI measures based on accuracy. These measures turn up in a variety of multisensory tasks, including detection, discrimination, recognition, and identification. We will not be able to cover all of them, but rather focus on a few important aspects.

2.5.1 MI Measures Based on Detection Accuracy

Let p_V, p_A, and p_{VA} denote the probability of responding "Yes" to the question of whether a visual, auditory, or combined visual–auditory stimulus has been presented, respectively. In analogy to CRE measures of response speed in the redundant signals task, we define the *crossmodal detection rate* as

$$\mathrm{CRE}_{DR} = \frac{p_{VA} - \max\{p_V, p_A\}}{\max\{p_V, p_A\}} \times 100. \tag{2.19}$$

Typically, the probability of a "Yes" response will primarily depend on stimulus intensity. If at least one of the unisensory stimuli is clearly detectable (i.e., p_A or p_V close to one), p_{VA} will also be close to one, and so the crossmodal detection rate will be close to zero. If intensity is low or, equivalently, the level of noise during presentation is (moderately) high, determining the likelihood of responding "Yes" is not straightforward: the participant may have a tendency to guess and/or may have an internal criterion for responding "Yes" or "No," which leads us to the realm of signal detection theory (SDT) (Green & Swets, 1974).

In the terminology of SDT, it is not sufficient to compare the crossmodal *hit rate* (probability of saying "Yes" when the stimulus is presented) with the unisensory

12 We limit the presentation here to SRTs; MI measures involving other aspects of eye movements are similarly obtainable.

hit rates because increasing the hit rate often goes along with increasing the *false-alarm rate* (probability of saying "Yes" when no stimulus is presented) as well. Assuming the standard equal-variance Gaussian distribution model of SDT, CRE_{DR} can be replaced by inserting the corresponding d-prime measures:

$$\mathrm{CRE}_{SDT} = \frac{d'_{VA} - \max\{d'_V, d'_A\}}{\max\{d'_V, d'_A\}} \times 100. \tag{2.20}$$

This measure assesses the relative amount of sensitivity increase in the visual–auditory condition compared to the best unisensory condition, while separating sensitivity from possible biases to respond "Yes" or "No" in each condition. An analogous definition for the focused attention task is obvious.

Measure CRE_{SDT} tests against a benchmark where the observer simply ignores the less detectable modality. However, it is also possible to modify CRE_{SDT} such that a PS strategy is taken into account. Let us assume that an observer sets two criteria, λ_V and λ_A, and a "Yes" response is given if at least one of the criteria is exceeded. Under stochastic independence, the probabilities of *misses* (1 minus probability of a hit) and *correct rejections* (1 minus probability of a false alarm) are the product of their modality components. Writing f_V, f_A, h_V, h_A and f_{VA}, h_{VA} for the false-alarm and hit rates for the unisensory and bisensory conditions, respectively, we get

$$\begin{aligned} f_{VA} &= 1 - (1 - f_V)(1 - f_A) = 1 - \Phi(\lambda_V)\Phi(\lambda_A) \\ h_{VA} &= 1 - (1 - h_V)(1 - h_A) = 1 - \Phi(\lambda_V - d'_V)\Phi(\lambda_A - d'_A), \end{aligned}$$

with Φ denoting the standard Gaussian distribution function. From this we can compute the visual–auditory sensitivity under the PS strategy:

$$d'^{PS}_{VA} = \Phi^{-1}(h_{VA}) - \Phi^{-1}(f_{VA}).$$

Inserting into expression (2.20) results in a modified measure of response enhancement gauging against PS:

$$\mathrm{CRE}^{PS}_{SDT} = \frac{d'_{VA} - d'^{PS}_{VA}}{d'^{PS}_{VA}} \times 100. \tag{2.21}$$

Besides the PS notion, numerous alternative models on how unisensory detection accuracy is combined into a bisensory one have been discussed in the literature (see Jones, 2016 for a recent tutorial). Finally, when there is empirical evidence against the equal-variance assumption of SDT, alternative measures, like the area under the operating characteristic, may be considered instead of d-prime values (see, e.g., Lovelace, Stein, & Wallace, 2003 for a focused-attention example).

2.5.2 Measures for Audiovisual Speech Identification

Arguably, one of the most thoroughly studied lines of multisensory research is the identification of speech in an audiovisual paradigm. In typical audiovisual speech

identification (or recognition) tests, listeners are presented with audio materials like syllables, words, phrases, or sentences along with a video of a speaker's face acquired at the same time as the audio materials. Commonly, speech heard in noise (often, talker babble noise at different levels) can be more accurately identified or recognized when the participant sees a speaker's articulating face or lip movements.

However, there still seems to be considerable controversy with respect to the source of this audiovisual advantage. According to several studies, when hearing-impaired individuals, or different age groups, are compared with respect to the amount of audiovisual benefit, one finds large differences across individuals or groups. Notably, these differences are often found to persist even when differing unisensory auditory or visual speech recognition performance levels are taken into account. Thus, besides lipreading ability and auditory encoding ability, an ability to integrate auditory and visual information should be assessed in order to explain audiovisual performance (Grant, 2002). In contrast, it is also held that an audiovisual speech signal represents a more robust representation of any given word because, first, simultaneous auditory and visual speech signals provide complementary information: vision contributes clues about some aspects of the speech event that are hard to hear and which may depend on the shape and contour of the lower face being clearly visible. Second, reinforcing information may be provided by the temporal congruence between amplitude fluctuations in the auditory signal and mouth opening and closing in the visual signal. That is, when the auditory signal gets louder, the visible mouth and jaw tend to be opening; when the signal gets softer, the mouth and jaw tend to be closing (see Tye-Murray *et al.*, 2016).

Measures of Response Enhancement and Superadditivity

Without subscribing to a specific source of the audiovisual advantage, ad-hoc measures of enhancement have been developed. Letting p_{AV} denote the probability[13] of correctly identifying words in the audiovisual condition and p_V,p_A the corresponding probability in the vision-only and auditory-only condition, respectively, one defines *visual enhancement* (VE) as

$$\mathrm{VE} = \frac{p_{AV} - p_A}{1 - p_A}. \tag{2.22}$$

Thus, VE represents the amount of benefit afforded by the addition of the visual channel of speech, normalized for the amount of possible improvement. Analogously, one defines *auditory enhancement* (AE) as

$$\mathrm{AE} = \frac{p_{AV} - p_V}{1 - p_V}. \tag{2.23}$$

Thus, AE represents the amount of benefit afforded by the addition of the auditory channel of speech, again normalized for the amount of possible improvement.

13 Note that p_V, p_A, and p_{AV} here are not the same as in the previous section on detection, but no confusion should arise.

Although these enhancement measures do not seem controversial, some criticism has been raised against them. First, whereas there is broad empirical support for the principle of inverse effectiveness (Section 2.3.1) being valid in audiovisual speech performance, the normalization involved in calculating AE biases against finding results consistent with it. Specifically, among listeners with equivalent improvement (i.e., equal numerators), AE will be lower for those who made more lipreading errors, inconsistent with the principle (as pointed out by Tye-Murray *et al.*, 2010, p. 639).

Second, a more sweeping argument was recently made by Dias, McClaskey, and Harris (2021), studying the mean proportion of correctly identified words for two different age groups. Consistent with previous research, they found p_V and p_A to decline with age and to correlate positively with each other, but p_{AV} did not differ significantly between age groups. Importantly, they did not find VE and AE to exhibit any age effects. Dias and colleagues offer the following explanation, after defining "superadditivity" p_{sAV} as

$$p_{sAV} = p_{AV} - (p_A + p_V). \tag{2.24}$$

Rewriting the expressions for VE and AE yields

$$\mathrm{VE} = \frac{p_{AV} - p_A}{1 - p_A} = \frac{p_V + p_{sAV}}{1 - p_A}$$

and

$$\mathrm{AE} = \frac{p_{AV} - p_V}{1 - p_V} = \frac{p_A + p_{sAV}}{1 - p_V}.$$

The superadditivity term occurring in both VE and AE explains the positive correlation; moreover, the authors argue, the absence of an age effect is due to the declining values of p_V and p_A with age, canceling an alleged increase of superadditivity, p_{sAV}, also with age.[14]

Measures Derived from Modeling Audiovisual Speech Identification

Different models of auditory–visual speech integration have been proposed. They often predict "optimal" performance in the bisensory condition given the information extracted in the unimodal conditions separately (e.g., for nonsense syllables, words, or sentences), thereby providing quantitative measures of *integration efficiency* (IE).

The simplest one is a model representing a PS version of crossmodal detection rate CRE_{DR} [Equation (2.19)]. Assuming independent PS for auditory and visual performance, the probability p^I_{AV} of recognizing an item in the audiovisual condition equals

$$p^I_{AV} = 1 - (1 - p_A) \times (1 - p_V) = p_A + p_V - p_A \times p_V.$$

14 Dias, McClaskey, and Harris (2021) use notation AO, VO, and AV instead of probabilities; see the paper for details of their exhaustive statistical analyses.

From this, integration efficiency is defined as (e.g., Tye-Murray, Sommers, & Spehar, 2007)

$$\mathrm{IE}^I = \frac{p_{AV}^{obs} - p_{AV}^I}{1 - p_{AV}^I}, \tag{2.25}$$

where p_{AV}^{obs} is the observed probability in the audiovisual condition. Integration efficiency measured this way has often been found to be positive, but some recent findings support the PS model as well (van de Rijt *et al.*, 2019), implying zero integration efficiency.

A prominent model for audiovisual speech identification is Massaro's *fuzzy logical model of perception* (FLMP), with an optimal integration rule equivalent to Bayes' theorem (see Massaro & Cohen, 2000).

Prelabeling model of integration (PRE). Another widely known model is Braida's PRE model (Braida, 1991), where each response R_j corresponds to a point in a D-dimensional Euclidean vector space of stimulus attributes (cue vectors) referred to as *prototypes*. Each presentation of a stimulus i generates a D-dimensional vector of cues X in the same space following a multivariate normal distribution with independent components, unit variance, and a given mean S_i not necessarily identical to the prototype corresponding to R_i. According to a decision rule of multidimensional signal detection theory, the subject responds R_j if and only if the (Euclidean) distance of X to the prototype of R_j is smaller than the distance to any other prototype. The prototype locations are assumed to reflect response bias effects, whereas the subject's sensitivity in discriminating stimulus i from stimulus j, d-prime value $d'(i,j)$, is given by the Euclidean distance between S_i and S_j. The model parameters (i.e., the components of vectors S_i and R_i) are estimated iteratively through nonmetric multidimensional scaling by comparing observed and predicted confusion matrices. The decision space for the AV condition is assumed to be the Cartesian product of the space for the A condition and the space for the V condition. A subject's sensitivity in the AV condition can be shown to be related to the unimodal sensitivities by

$$d'_{AV}(i,j) = \sqrt{d'_A(i,j)^2 + d'_V(i,j)^2}. \tag{2.26}$$

An IE measure is then defined by taking the ratio between the obtained and predicted d'_{AV} scores:

$$\mathrm{IE}^{PRE} = \frac{d'_{AV}(obs)}{d'_{AV}(pred)}. \tag{2.27}$$

Note that perfect integration need not be associated with high overall AV performance: if a participant has very bad hearing or is a very poor speech reader, it is unlikely that they will achieve a high AV score. Nevertheless, a subject may still integrate the available A and V cues in a nearly optimal manner, and if so, the integration efficiency measure should be near unity (see Figure 2.3).

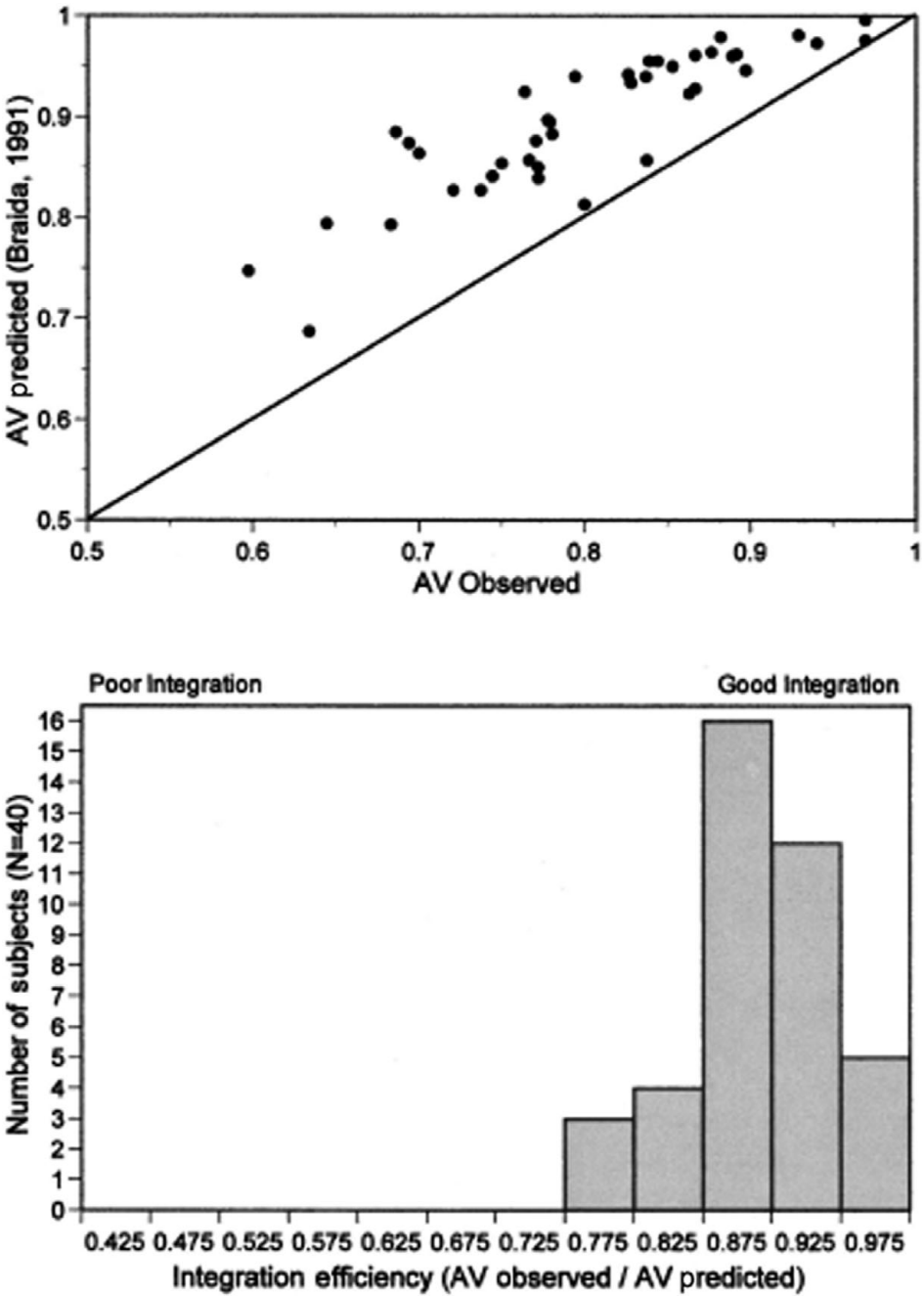


Figure 2.3 *PRE model: Observed and derived measures obtained from experiment on consonant recognition in noise (40 subjects). (Top) Observed vs. predicted PRE AV scores. The line indicates perfect integration efficiency:* $\mathrm{IE}^{PRE} = 1$. *Predicted and observed AV scores for several subjects fall near the main diagonal, whereas observed scores for other subjects are significantly less than predicted. (Bottom) Histogram showing distribution of* IE^{PRE} *values across subjects (from Grant & Seitz, 1998).*

Integration Efficiency Based on Fechnerian Scaling

The validity of any IE measure derived from a model of AV speech integration, like the prelabeling model (PRE), depends on the specific assumptions of the model being valid empirically. We briefly discuss an alternative, less restrictive approach based on a theory of computing *subjective distances* on very general stimulus sets (Dzhafarov & Colonius, 2006).

Recall that a *metric* is a non-negative function d defined on pairs (x,y) from a set X, say, such that for all $x,y,z \in X$:

(i) $d(x,y) \geq 0$ and $d(x,y) = 0$ implies $x = y$;
(ii) $d(x,y) = d(y,x)$;
(iii) $d(x,y) + d(y,z) \geq d(x,z)$.

The theory of *Fechnerian scaling* (FS) (see, e.g., Dzhafarov & Colonius, 2007) deals with the computation of subjective distances among stimuli from their pairwise discrimination probabilities. The latter are the probabilities with which the judgment "these two stimuli are different" is chosen over "these two stimuli are the same":

$$\psi(x,y) = \mathrm{P}\,[\text{subject judges } x \text{ and } y \text{ in } (x,y) \text{ to be different}]. \tag{2.28}$$

For identification tasks, data from confusion matrices are available instead of discrimination probabilities. The cell in a confusion matrix is the probability that stimulus y is identified as stimulus x, denoted as $\eta(x,y)$ for all x,y in the stimulus set X. Thus, we need the additional assumption that

$$1 - \psi(x,y) = \eta(x,y).$$

Given $\eta(x,y)$ for all x,y in the stimulus set X, FS allows one to compute a metric G, say, on X satisfying properties (i) to (iii) above. The only necessary and sufficient empirical condition for the construction is *regular maximality*:

$$\eta(x,x) > \max\{\eta(x,y), \eta(y,x)\} \tag{2.29}$$

for any $x,y \in X, x \neq y$. In other words, when stimulus x is presented, the probability of identifying x as x should be greater than the probability of identifying x as y, a stimulus different from x. Importantly, $\eta(x,x)$ may vary with x and $\eta(x,y)$ may be different from $\eta(y,x)$.

Let us assume that Fechnerian metrics G_A, G_V, and G_{AV} have been computed from the confusion matrices in the auditory, visual, and audiovisual condition, respectively, for each pair of stimuli $\{i,j\}$. The corresponding metric values $G_A(i,j)$, $G_V(i,j)$, and $G_{AV}(i,j)$ are interpreted as subjective distance between the two stimuli under auditory, visual, and audiovisual presentation, respectively. A priori, these three values are unrelated to each other since they are defined on different stimulus sets. On the other hand, there is a natural one-to-one correspondence across the visual, auditory, and bisensory stimulus sets (i.e., visual stimulus $i \leftrightarrow$ auditory stimulus $i \leftrightarrow$ bisensory stimulus component i). Moreover, given that Fechnerian distances on a given set are unique only up to a similarity transformation

(i.e., multiplication with a positive constant), one can standardize each of them such that the maximum distance equals one.[15]

If $G_{AV}(i,j)$ is larger than $G_A(i,j)$ or $G_V(i,j)$, this suggests that adding information from the other modality (V or A) increases the subjective distance between i and j. This increase in subjective distance from the unisensory to the bisensory presentation is proposed as indicator of visual, respectively auditory enhancement, in analogy to VE and VA in Section 2.5.2:

$$\begin{aligned} \mathrm{VE}^{FS}(i,j) &= \frac{G_{AV}(i,j) - G_A(i,j)}{1 - G_A(i,j)} \\ &= \frac{G_V(i,j) + G_{sAV}(i,j)}{1 - G_A(i,j)} \end{aligned} \tag{2.30}$$

and

$$\begin{aligned} \mathrm{VE}^{FS}(i,j) &= \frac{G_{AV}(i,j) - G_V(i,j)}{1 - G_V(i,j)} \\ &= \frac{G_A(i,j) + G_{sAV}(i,j)}{1 - G_V(i,j)}, \end{aligned} \tag{2.31}$$

with

$$G_{sAV}(i,j) = G_{AV}(i,j) - [G_A(i,j) + G_V(i,j)]$$

denoting the superadditivity term, in analogy to Equation (2.24).

In order to derive an overall index of integration efficiency, averaging across all superadditivity terms results in a *Fechnerian scaling-based multisensory integration efficiency* index:

$$\mathrm{IE}^{FS} = \binom{N}{2}^{-1} \sum_{\{i,j\} \subset S} G_{sAV}(i,j), \tag{2.32}$$

$i \neq j$, with N denoting the number of stimuli in stimulus set S.

The FS-based approach to integration efficiency presented here, and the prelabeling model (PRE), share the idea of converting the information contained in the confusion matrices into a representation of subjective distances between the stimuli. An important difference is that the FS-based approach neither requires explicit assumptions about the space (e.g., Euclidean) and its dimensionality nor any parameter estimation.

One can argue that the definition of IE^{FS} being based on superadditivity is somewhat arbitrary. Nonetheless, Colonius and Diederich (2007) report on a small data set, a reduced confusion matrix for consonants /b/, /d/, and /g/ presented in Braida, Sekiyama, and Dix (1998). Table 2.2 lists all three confusion matrices

15 Importantly, Fechnerian distances are always a function of the entire (base) set used to compute them, and the G values are not monotonically related to the probabilities $\eta(x,y)$, although they have been found to correlate highly in many empirical data sets. Moreover, it seems plausible that Fechnerian distances for corresponding stimulus pairs are measured in the same "units."

Table 2.2 *Each cell: ψ at top and G (Fechnerian distances) at bottom, for auditory (A), visual (V), and audiovisual (AV) presentation (rows ≡ stimuli, columns ≡ responses) with resulting value of* $IE^{FS} = 0.8737$.

$\psi_A = 1 - \eta_A$ G_A	"b"	"d"	"g"
	0.437	0.717	0.846
b–	0.000	0.450	0.589
	0.700	0.530	0.757
d–	0.450	0.000	0.350
	0.746	0.689	0.566
g–	0.589	0.350	0.000
$\psi_V = 1 - \eta_V$ G_V			
	0.022	0.983	0.996
–b	0.000	1.805	1.527
	0.990	0.146	0.871
–d	1.805	0.000	0.864
	0.989	0.575	0.436
–g	1.527	0.864	0.000
$\psi_{AV} = 1 - \eta_{AV}$ G_{AV}			
	0.007	0.996	0.998
bb	0.000	1.860	1.704
	0.997	0.126	0.876
dd	1.860	0.000	1.203
	0.991	0.731	0.278
gg	1.704	1.203	0.000

(auditory, visual, auditory–visual) together with their corresponding Fechnerian distances G_A, G_V, and G_{VA}.

The value of IE^{FS} was computed[16] as 0.8737, which is very close to the correct identification score (87.1%) predicted by the PRE model (Braida, Sekiyama, & Dix, 1998) for the same data set. In general, however, most of the indexes of audiovisual integration efficiency presented here have some degree of arbitrariness and will have to prove their utility and cross-study consistency in future research.

16 The IE^{FS} index used was based on the superadditivity term $G_{sAV}(i,j)$ written as ratio rather than difference.

2.6 Measures Based on MI Modeling of RTs

The focus of this chapter has so far been on measuring MI, rather than modeling. Yet PS, which is a model, has emerged several times as benchmark: any improvement (response speed reduction, improved detection probability, etc.) beyond the level predicted by PS has been defined as measure of MI. In keeping with this approach, we will define CRE as a function of the enhancement observed beyond what is predicted by a particular MI model under consideration. Given that these models typically require estimation of some parameters, the idea here is to estimate them from the unisensory conditions only and subsequently insert these estimates into the MI model in order to predict bisensory RTs. Measures of MI then assess by how much these model predictions fall short of the observed bisensory data. Given the multitude of integration models, however, we need to be selective and will only sketch a few modeling approaches with respect to how they estimate and predict the amount of MI.

2.6.1 Coactivation Models

Coactivation is a generic term suggested by J. Miller (1982) to describe models that allow activation from different channels (in particular, modalities) to combine in satisfying a single criterion for response initiation, in distinction to *separate activation* models (or, race models), where the system never combines activation from different channels in order to meet its criterion for responding (J. Miller, 1982, p. 248). Coactivation models differ with respect to their state space (i.e., whether the state space within which combination is performed, is continuous or discrete). We consider measures of MI for continuous-time models with either discrete or continuous state space. Discrete-time coactivation models are not considered here because of our emphasis on response time measurements.

The (Poisson) superposition model. Presentation of a stimulus induces a neural renewal (counting) process,[17] $\{N(t), t \geq 0\}$, with interarrival times $\{X_n, n = 1, 2\ldots\}$. Let $W(n) = \sum_{i=1}^{n} X_i$ be the waiting time for the nth counts. The assumption is that a response is initiated as soon as a fixed number of counts, c, is reached. Note that

$$P(N(t) \geq c) = P(W(c) \leq t).$$

Finally, the observable RT is assumed to be additively composed of the waiting time plus all processes following (or preceding) it. The duration of these additional processes, which may include motor preparation and response execution components, is represented by a random variable M:

$$RT = W(c) + M.$$

17 For exact definitions, see Ross (1996).

The superposition assumption holds that the unisensory renewal processes, $N_V(t)$ and $N_A(t)$, are simply added, defining a new renewal process, so that the waiting time for the cth count is reduced; specifically, if the visual stimulus is presented τ ms ($\tau > 0$) before the auditory:

$$N_{VA}(t) = N_V(t) + N_A(t - \tau),$$

where $N_A(t - \tau) = 0$ for $t < \tau$.

Under the simplest renewal process (Poisson), expected waiting time for the bisensory condition can be computed as

$$\mathrm{E}W_{V\tau A}(c) = \frac{c}{\alpha_V} - \frac{\alpha_A}{\alpha_V(\alpha_V + \alpha_A)} \exp(-\alpha_V \tau) \sum_{i=0}^{c-1} \frac{(\alpha_V \tau)^i}{i!} (c - i), \tag{2.33}$$

where α_V and α_A are the Poisson intensity parameters for the visual and auditory stimulus, respectively.[18] For $\tau = 0$, this reduces to $c/(\alpha_V + \alpha_A)$.

Let $\mathrm{E}RT_{V\tau A} = \mathrm{E}W_{V\tau A}(c) + \mathrm{E}M$. In obvious notation, we define as measure of crossmodal response enhancement for the Poisson superposition model:

$$\mathrm{CRE}_{SUP,\tau} = \frac{\mathrm{E}RT_{V\tau A} - \mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100, \tag{2.34}$$

assuming parameters c and EM to be invariant across the unisensory and bisensory conditions. Note that $\mathrm{CRE}_{SUP,\tau}$ increases as a function of c; thus, the Poisson superposition model is consistent with the prediction of inverse effectiveness. On the other hand, it cannot predict inhibition.

Diffusion models. In these models, presentation of a stimulus is assumed to induce a stochastic process that is often described by a linear, first-order stochastic differential equation[19] of the form

$$dX(t) = \mu(X(t), t) + \sigma(X(t), t)\, dW(t), \tag{2.35}$$

where $W(t)$ is a standard *Wiener process*, $\mu(x,t)$ is called the *effective drift rate* describing the instantaneous rate of expected increment change at time t and state $x = X(t)$. Factor $\sigma(X(t), t)$ in front of the instantaneous increments $dW(t)$ is called the *diffusion coefficient* relating to the variance of the increments.

Modeling information accumulation and predicting response times, however, requires one to make concrete assumptions on drift rates and diffusion coefficients, resulting in a large variety of stochastic diffusion models. For example, setting $\mu(x,t) = \delta$ and $\sigma(X(t), t) = \sigma$ defines a time-homogeneous Wiener process with drift (setting $\delta = 0$ is the standard Wiener process). The drift rate is interpreted as describing the rate of information accumulation under different stimulus conditions.

Termination of the accumulation process is then defined by the first time it reaches a threshold, C ($C > 0$). This *stopping time*, denoted as ν, is the smallest

18 For $\tau < 0$, τ must be replaced by $-\tau$ and α_V and α_A interchanged.

19 For exact definitions, we must refer to the literature.

value for t such that $X(t) = C$. If $X(0) = 0$, expected stopping time in the Wiener process with drift δ is

$$\mathrm{E}[\nu \mid X(0) = 0] = C/\delta, \tag{2.36}$$

which is independent of the diffusion coefficient. Observed RT is defined as the sum of (random variables) ν and a non-decision component M, $RT = \nu + M$.

Applying this model version to the redundant signals paradigm, we assume two Wiener processes with drift rates δ_V and δ_A, respectively, for the unisensory conditions. In the bisensory condition with $SOA \equiv \tau = 0$, a superposed Wiener process is defined by

$$X_{VA}(t) = X_V(t) + X_A(t) \tag{2.37}$$

with drift rate $\delta_V + \delta_A$, while postulating identical threshold values C and mean values of M, for all conditions. Given the expected stopping times $C/\delta_V, C/\delta_A, C/(\delta_V + \delta_A)$, one can define a measure of crossmodal enhancement exactly like Equation (2.34) for the Poisson superposition model at $\tau = 0$. Obviously, however, under these simplified assumptions the two models become indistinguishable, predicting the same amount of enhancement. The problem dissolves when predictions for non-simultaneous stimuli for two modalities (Schwarz, 1994) or more (Diederich, 1995) are derived and CRE measures analogous to Equation (2.34) can be defined:

$$\mathrm{CRE}_{DIF,\tau} = \frac{\mathrm{E}RT_{V\tau A} - \mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100. \tag{2.38}$$

Moreover, for $\tau = 0$, setting $\mu(x,t) = \delta - \gamma x$ and $\sigma(x,t) = \sigma$ defines a time-homogeneous *Ornstein–Uhlenbeck process* (OUP).[20] For $\gamma > 0$ this implies that the accumulation rate decays depending on the current state x (e.g., Diederich, 1995). Given that for this and related models, expected stopping times are often not available in closed form, crossmodal enhancement measures of the form of Equation (2.38) may be approximated by simulation or, alternatively, by Markov chain approximation.[21]

2.6.2 Time-Window-of-Integration Framework

While the PS mechanism by itself constitutes a broad class of models at both the neural and behavioral level, simple race models often do not fare too well empirically and, as mentioned, typically only serve as point of reference in defining an enhancement measure (see Section 2.4.3). The time-window-of-integration (TWIN) framework for response speed, measured as manual or saccadic RT, is

20 But $\tau \neq 0$ implies a non-time-homogeneous OU process.

21 Roughly, after fitting the unisensory data with an OUP model each, sample unisensory values $x_V(t)$ and $x_A(t)$ for any t, add them to define a superposed process, and estimate the expected stopping time of that process.

a simple extension of the PS model. The amount of RT facilitation not accounted for by the latter [cf. Equation (2.17)] is,

$$\mathrm{E}\min\{RT_V, RT_A\} - \mathrm{E}RT_{VA},$$

where RT_A and RT_V are the observed latencies of unisensory responses. Here, $\mathrm{E}\min\{RT_V, RT_A\}$ refers to the RT predicted by a PS rule (stochastically independent or dependent race) and $\mathrm{E}RT_{VA}$ is the observed bisensory mean RT. The TWIN framework postulates two serial processing stages. A first (race) stage among the activity elicited by the different modalities is followed by a second stage that is defined by default: it includes all subsequent, possibly temporally overlapping, processes that are not part of the processes in the first stage, and crossmodal interaction can only occur in the second stage.

While the framework is mute about the specific mechanism of integration in the second stage, its central feature is the notion of a time-window of MI. It postulates that crossmodal interaction occurs *only* if the peripheral processes of the first stage all terminate within a given temporal interval, the "time window of integration." The result of crossmodal interaction manifests itself in an increase, or decrease, of second-stage processing time. The window acts as a filter determining whether afferent information delivered from different sensory organs is registered close enough in time to trigger MI. Passing the filter is necessary, but not sufficient, for crossmodal interaction to occur, because the amount of interaction may also depend on many other aspects of the stimulus context, in particular the spatial configuration of the stimuli.[22] Although the amount of interaction does not depend directly on stimulus onset asynchrony (SOA) of the stimuli, temporal tuning of the interaction still occurs because the *probability of the integration event* is modulated by the SOA value. Formalization of the framework makes these observations explicit.

We introduce some notation and derive an expression for the measure of MI in the TWIN framework. With τ ($-\infty < \tau < +\infty$) as SOA value and ω ($\omega \geq 0$) as parameter for the integration window width, these assumptions imply that the event that MI occurs, denoted by I, equals

$$\begin{aligned} I &\equiv \{|T_V - (T_A + \tau)| < \omega\} \\ &= \{T_A + \tau < T_V < T_A + \tau + \omega\} \cup \{T_V < T_A + \tau < T_V + \omega\}, \end{aligned}$$

where T_V, T_A are assumed to be continuous random variables and the presentation of the visual stimulus is (arbitrarily) defined as the physical zero time point. Thus, the probability of integration occuring, $P(I)$, is an increasing function of ω, but its dependence on τ will be a function of the specific distributions assumed for T_V and T_A.

22 Note that the window of the TWIN framework is only defined temporally, in contrast to the spatio-temporal window sometimes postulated.

Writing S_1 and S_2 for first and second-stage processing times, respectively, overall expected RT in the crossmodal condition with an SOA equal to τ, $\mathrm{E}[RT_{V\tau A}]$, is computed conditioning on event I (integration) occurring or not:

$$\begin{aligned}\mathrm{E}[RT_{V\tau A}] &= \mathrm{E}[S_1] + P(I)\,\mathrm{E}[S_2|I] + [1 - P(I)]\,\mathrm{E}[S_2|I^c] \\ &= \mathrm{E}[S_1] + \mathrm{E}[S_2|I^c] - P(I) \times \Delta. \\ &= \mathrm{E}[\min(T_V, T_A + \tau)] + \mathrm{E}[S_2|I^c] - P(I) \times \Delta. \end{aligned} \tag{2.39}$$

Here, I^c denotes the event complementary to I and Δ stands for $\mathrm{E}[S_2|I^c] - \mathrm{E}[S_2|I]$. The term $P(I) \times \Delta$ is a measure of the expected amount of crossmodal interaction in the second stage, with positive Δ values corresponding to facilitation, negative ones to inhibition. Because event I cannot occur in the unimodal (visual or auditory) condition, expected RT under these conditions is, respectively:

$$\mathrm{E}[RT_V] = \mathrm{E}[T_V] + \mathrm{E}[S_2|I^c] \quad \text{and} \quad \mathrm{E}[RT_A] = \mathrm{E}[T_A] + \mathrm{E}[S_2|I^c].$$

Note that the race in the first stage produces a not directly observable statistical facilitation effect (*SFE*) analogous to the one in the "classic" race model:

$$SFE \equiv \min\{\mathrm{E}[T_V], \mathrm{E}[T_A] + \tau\} - \mathrm{E}[\min\{T_V, T_A + \tau\}].$$

This contributes to the overall crossmodal interaction effect predicted by TWIN, which amounts to

$$\min\{\mathrm{E}[RT_V], \mathrm{E}[RT_A] + \tau\} - \mathrm{E}[RT_{V\tau A}] = SFE + P(I) \times \Delta.$$

Thus, in the TWIN framework crossmodal *facilitation* observed in a redundant signals task may be due to MI or statistical facilitation, or both. This shows that the TWIN extends the race model class by predicting integration effects over and above statistical facilitation. Moreover, a potential multisensory *inhibitory* effect occurring in the second stage may be weakened, or even masked completely, by the simultaneous presence of statistical facilitation in the first stage.

We have shown that one can derive various empirically testable predictions from the TWIN framework even without assuming specific distributions for the random processing times. In addition, when T_V and T_A are independent and exponentially distributed random variables and the expected value for second-stage processing time with no crossmodal interaction is set as parameter μ, then numerical estimates of the overall crossmodal interaction effect, $SFE + P(I) \times \Delta$, are available. This suggests the following definition for crossmodal enhancement:

$$\mathrm{CRE}_{TWIN} = \frac{\mathrm{E}RT_{V\tau A} - \mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100, \tag{2.40}$$

with $\mathrm{E}RT^{obs}_{V\tau A}$ denoting observed mean bisensory RT and $\mathrm{E}RT_{V\tau A}$ the expected bisensory RT under the TWIN model, which can be calculated using parameter estimates obtained from fitting the model to the observations.

Note that "temporal window of integration" has become an important concept in describing crossmodal binding effects as a function (e.g., of age, specific disorders,

or training in a variety of MI tasks apart from RTs).[23] In fact, the width of the time window can by itself be taken as a measure for MI: in the temporal order judgment (TOJ) task, where subjects are required to judge the order of stimuli (visual first vs. auditory first), the width of the window determines how often the two stimuli will be "bound together" and, thereby, how often the subject can only guess that the visual stimulus occurred first. Within a simple extension of the TWIN framework to include the TOJ task, widening the temporal window of integration in a RT task, or narrowing it in a TOJ task, can be seen as an observer's strategy to optimize performance in an environment where the temporal structure of sensory information from separate modalities provides a critical cue for inferring the occurrence of crossmodal events (Diederich & Colonius, 2015).

2.7 Conclusions

It turned out that, in order to construct valid measures of integration, a possible effect of PS had to be taken into account, in both behavioral and neural contexts. Specifically, we have argued that the common index for RTs in the redundant signals paradigm [see Equation (2.2)],

$$\mathrm{CRE}_{RT} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\} - \mathrm{E}RT_{VA}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\}} \times 100, \quad (2.2)$$

should be replaced by assuming a race model with maximal negative dependence:

$$\mathrm{CRE}_{RT}^{min} = \frac{\mathrm{E}^{(-)} \min\{RT_V, RT_A\} - \mathrm{E}RT_{VA}}{\mathrm{E}^{(-)} \min\{RT_V, RT_A\}} \times 100,$$

which is Equation (2.17) for $\tau = 0$. The latter is a more conservative index because it allows for the possibility that the "race" between visual and auditory activation may be (maximally) negatively dependent in the statistical sense, that is, it measures how much faster observed mean RT is than the fastest one that can be generated by PS alone. See Table 2.3 for a list of all indexes used in the chapter.

A further argument in favor of using $\mathrm{CRE}_{RT}^{(-)}$ is that $\mathrm{E}^{(-)} \min\{RT_V, RT_A\}$ can be sensitive to the shape of the entire distribution of the unisensory RT distributions, like moments higher than the mean, see Colonius and Diederich (2017). Another, non-RT, example is a discrimination task where estimator variance is required to obtain a statistically optimal linear combination of modalities (Drugowitsch *et al.*, 2014; Ernst & Banks, 2002), so that any MI measure gauging the degree of deviation from optimality will be a function of the second moment.

Thus, instead of defining MI measures via means only, it may be argued that one should compare entire distributions in order to obtain more informative measures.

23 It is worth pointing out that the time window concept in the TWIN framework differs from the one used in most empirical studies. The latter is typically defined by the range of SOA values wherein crossmodal effects can be observed. In contrast, in the former (i) window width is a parameter to be estimated from the data, and (ii) the filter is not in principle limited to the temporal structure of the stimulus context but could be defined more broadly (e.g., including spatial features or subjective values; see Bean, Stein, & Rowland, 2021).

Table 2.3 *List of all indexes in the chapter (for spikes, RTs, detection accuracy, AV speech identification, RT models).*

Type	Index	Definition	Section
spikes	CRE_{SP}	$\frac{\mathrm{E}N_{VA}-\max\{\mathrm{E}N_V,\mathrm{E}N_A\}}{\max\{\mathrm{E}N_V,\mathrm{E}N_A\}}\times 100$	2.3.3
	CRE^{*}_{SP}	$\frac{\mathrm{E}N_{VA}-\mathrm{E}\max\{N_V,N_A\}}{\mathrm{E}\max\{N_V,N_A\}}\times 100$	2.3.3
	CRE^{max}_{SP}	$\frac{\mathrm{E}N_{VA}-\mathrm{E}^{(-)}\max\{N_V,N_A\}}{\mathrm{E}^{(-)}\max\{N_V,N_A\}}\times 100$	2.3.3
RTs	$\mathrm{CRE}_{RT,\tau}$	$\frac{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau\}-\mathrm{E}RT_{V\tau A}}{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau\}}\times 100$	2.4.1
	$\mathrm{CRE}_{RT,\tau_1\tau_2}$	$\frac{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau_1,\mathrm{E}RT_T+\tau_1+\tau_2\}-\mathrm{E}RT_{V\tau_1A\tau_2T}}{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau_1,\mathrm{E}RT_T+\tau_1+\tau_2\}}\times 100$	2.4.1
	$\mathrm{CRE}_{RT,(\tau_1)\tau_2}$	$\frac{\mathrm{E}RT_{V\tau_1A}-\mathrm{E}RT_{V\tau_1A\tau_2T}}{\mathrm{E}RT_{V\tau_1A}}\times 100$	2.4.1
	CRE^{min}_{RT}	$\frac{\mathrm{E}^{(-)}\min\{RT_V,RT_A\}-\mathrm{E}RT_{VA}}{\mathrm{E}^{(-)}\min\{RT_V,RT_A\}}\times 100$	2.4.1
accuracy	CRE_{DR}	$\frac{p_{VA}-\max\{p_V,p_A\}}{\max\{p_V,p_A\}}\times 100$	2.5.1
	CRE_{SDT}	$\frac{d'_{VA}-\max\{d'_V,d'_A\}}{\max\{d'_V,d'_A\}}\times 100$	2.5.1
	CRE^{PS}_{SDT}	$\frac{d'_{VA}-d'^{PS}_{VA}}{d'^{PS}_{VA}}\times 100$	2.5.1
AV speech	p_{sAV} (superadditivity)	$p_{AV}-(p_A+p_V)$	2.5.2
	VE (vis. enhancement)	$\frac{p_{AV}-p_A}{1-p_A}=\frac{p_V+p_{sAV}}{1-p_A}$	2.5.2
	AE (aud. enhancement)	$\frac{p_{AV}-p_V}{1-p_V}=\frac{p_A+p_{sAV}}{1-p_V}$	2.5.2
	p^{I}_{AV}	$1-(1-p_A)\times(1-p_V)$	2.5.2
	IE^{I} (integr. efficiency)	$\frac{p^{obs}_{AV}-p^{I}_{AV}}{1-p^{I}_{AV}}$	2.5.2
	IE^{PRE}	$d'_{AV}(obs)/d'_{AV}(pred)$	2.5.2
	$G_{sAV}(i,j)$	$G_{AV}(i,j)-[G_A(i,j)+G_V(i,j)]$	2.5.2
	IE^{FS}	$\binom{N}{2}^{-1}\sum_{\{i,j\}\subset S}G_{sAV}(i,j)$	2.5.2
RT model	$\mathrm{CRE}_{SUP,\tau}$	$\frac{\mathrm{E}RT_{V\tau A}-\mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}}\times 100$	2.6.1
	$\mathrm{CRE}_{DIF,\tau}$	$\frac{\mathrm{E}RT_{V\tau A}-\mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}}\times 100$	2.6.1
	CRE_{TWIN}	$\frac{\mathrm{E}RT_{V\tau A}-\mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}}\times 100$	2.6.2

Assume there exists a numerical function δ measuring the distance between two distributions (e.g., $\delta(F_A,F_{VA})$); one may define crossmodal response enhancement, in analogy to CRE_{RT} above, by

$$\mathrm{CRE}_{\delta}=\min\{\delta(F_V,F_{VA}),\delta(F_A,F_{VA})\}\times 100. \tag{2.41}$$

Here, δ is already normalized to a range from zero to one; if F_{VA} is equal to one of the unisensory distributions, then $\mathrm{CRE}_\delta = 0$.[24] Thus, the first two requirements for a CRE measure (see Section 2.2.2) are satisfied, while the inhibition case is not covered. More complex measures are certainly possible; however, a more pressing task is to find criteria for selecting some measure δ from the "universe" of distance measures between distributions that would make the choice less arbitrary.

2.8 Related Literature

Despite the limited scope of this chapter, we hope to have given a first glimpse into the various ways and issues of defining measures of MI. A broader and deeper view may be gained from the references given in this section.

A number of comprehensive handbooks and review articles on MI are available: Bremner, Lewkowicz, and Spence (2012); Calvert, Spence, and Stein (2004); Colonius and Diederich (2020); Murray and Wallace (2012); Naumer and Kayser (2010); B. E. Stein (2012); Stevenson *et al.* (2014); van Opstal (2016). The first monograph on MI from the neurophysiology point of view is B. E. Stein and Meredith (1993), while B. E. Stein *et al.* (2009) discuss quantitative methods for measuring MI at the single-neuron level. Early studies by Todd (1912), measuring RT to stimuli from two or more sensory modalities, presented both singly and together, are often seen as the beginnings of the scientific study of crossmodal behavior. Raab (1962) is the classic reference for a treatment of the "race model" and PS mechanisms for RTs. The latter has typically been presented under the hypothesis of stochastic independence. The "race model inequality" [see Equation (2.15)], first developed by J. Miller (1982) and tested by Diederich and Colonius (1987), initiated the discussion of non-independent PS in the context of copula theory (Colonius, 1990, 2016; Colonius & Diederich, 2017) and the development of related statistical tests (Gondan, 2010; Gondan, Riehl, & Blurton, 2012; Lombardi, D'Allesandro, & Colonius, 2019; Ulrich, Miller, & Schröter, 2007). Generalized race model inequalities have been discussed in Colonius, Woff, and Diederich (2017); Gondan, Dupont, and Blurton (2020); Gondan and Vorberg (2021). The "principle of congruent effectiveness" (Otto, Dassy, & Mamassian, 2013), stating that multisensory behavior (specifically, speedup of response times) is largest when behavioral performance in corresponding unisensory conditions is similar, corresponds to the index of unisensory imbalance (UI) [see Equation (2.3)].

Regarding accuracy measures, Jones (2016) provides a comprehensive tutorial about models of cue combination based on measures of sensitivity, including signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002). Schwarz and Miller (2014) point out that PS does not always lead to facilitation in compound detection and discrimination tasks because an increase of hit rate may also cause an increase of false alarms; evaluating unisensory vs. bisensory performance should,

24 Probability summation could be accounted for by defining $\mathrm{CRE}_\delta = \delta(\min\{F_V(t)+F_A(t),1\}, F_{VA}) \times 100$.

therefore, be performed via comparing the associated areas under the ROC curves. Billock *et al.* (2021) present a framework for comparing spike rates from AV integration in cortical bisensory neurons with psychophysical (discrimination) data and suggest vector-like Minkowski combination models describing either.

The literature on AV speech processing is huge; the handbook by Bailly, Perrier, and Vatikiotis-Bateson (2012) is a good source, as well as reports from the *International Conference on Auditory-Visual Speech Processing* (AVSP).[25] More details on the Fechnerian scaling approach can be found in the chapter by E. N. Dzhafarov and H. Colonius in this volume (*Fechnerian Scaling: Dissimilarity Cumulation Theory*).

The Poisson superposition model for MI has been introduced in Schwarz (1989) and discussed in Diederich (1995), Diederich and Colonius (1991), and Schwarz (1994). A tutorial on diffusion processes for RTs is given in Smith (2000), and a comprehensive treatment of stochastic models for decision-making is the chapter by Diederich and Mallahi-Karai in Volume II (Diederich & Mallahi-Karai, 2018). Notably, diffusion models can be extended to describe binary choice response tasks by assuming an upper and a lower absorbing bound for the accumulation process (Ratcliff, 1978). Such a diffusion superposition model for audiovisual data is discussed and tested by experiment in Blurton, Greenlee, and Gondan (2014). Drugowitsch *et al.* (2014) introduce a diffusion model for visual–vestibular integration with a weighted superposition approach that accumulates evidence optimally across both cues and time. For other extensions of diffusion models, see Diederich (1997), Diederich and Oswald (2016), and Mallahi-Karai and Diederich (2021). The time-window-of-integration model was introduced by the authors in 2004 (Colonius & Diederich, 2004) and subsequently extended and experimentally tested in Diederich and Colonius (2007a, 2007b, 2008a, 2008b).

References

Bailly, G., Perrier, P., & Vatikiotis-Bateson, E. (Eds.). (2012). *Audiovisual speech processing*. Cambridge: Cambridge University Press.

Bean, N., Stein, B., & Rowland, B. (2021). Stimulus value gates multisensory integration. *European Journal of Neuroscience*, *53*, 3142–3159.

Billock, V., Kinney, M., Schnupp, J., & Meredith, M. (2021). A simple vector-like law for perceptual information combination is also followed by a class of cortical multisensory bimodal neurons. *iScience*, https://doi.org/10.1016/j.isci.2021.102527.

Blurton, S., Greenlee, M., & Gondan, M. (2014). Multisensory processing of redundant information in go/no–go and choice responses. *Attention, Perception, and Psychophysics*, *76*, 1212–1233.

Braida, L. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology A*, *43*(3), 647–677.

25 www.isca-speech.org/archive

Braida, L., Sekiyama, K., & Dix, A. (1998). Integration of audiovisually compatible and incompatible consonants in identification experiments. In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Eds.), *Auditory-visual speech processing 1998 (avsp98)* (pp. 49–54).

Bremner, A., Lewkowicz, D., & Spence, C. (Eds.). (2012). *Multisensory development*. Oxford: Oxford University Press.

Calvert, G., Spence, C., & Stein, B. E. (2004). *Handbook of multisensory processes*. Cambridge, MA: MIT Press.

Colonius, H. (1990). Possibly dependent probability summation of reaction time. *Journal of Mathematical Psychology*, *34*, 253–275.

Colonius, H. (2016). An invitation to coupling and copulas, with applications to multisensory modeling. *Journal of Mathematical Psychology,*, *74*, 2–10.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: A time-window-of-integration model. *Journal of Cognitive Neuroscience*, *16*, 1000–1009.

Colonius, H., & Diederich, A. (2007). A measure of auditory-visual integration efficiency based on fechnerian scaling. In J. Vroomen, M. Swerts, & E. Krahmer (Eds.), *Auditory-visual speech processing 2007 (avsp2007).*

Colonius, H., & Diederich, A. (2017). Measuring multisensory integration: From reaction times to spike counts. *Scientific Reports*, *7*(1), 3023. (http://dx.doi.org/10.1038/s41598-017-03219-5)

Colonius, H., & Diederich, A. (2020). Formal models and quantitative measures of multisensory integration: A selective overview. *European Journal of Neuroscience*, *51*, 1161–1178.

Colonius, H., Diederich, A., & Steenken, R. (2009). Time-window-of-integration (twin) model for saccadic reaction time: Effect of auditory masker level on visual-auditory spatial interaction in elevation. *Brain Topography*, *21*, 177–184.

Colonius, H., Woff, F., & Diederich, A. (2017). Trimodal race model inequalities in multisensory integration. *Frontiers in Psychology*, *8*(1141).

Dias, J., McClaskey, C., & Harris, K. (2021). Audiovisual speech is more than the sum of its parts: Auditory-visual superadditivity compensates for age-related declines in audible and lipread speech intelligibility. *Psychology and Aging*, *36*(4), 520–530.

Diederich, A. (1995). Intersensory facilitation of reaction time: Evaluation of counter and diffusion coactivation models. *Journal of Mathematical Psychology*, *39*, 197–215.

Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, *41*, 260–274.

Diederich, A., & Colonius, H. (1987). Intersensory facilitation in the motor component? *Psychological Research*, *49*, 23–29.

Diederich, A., & Colonius, H. (1991). A further test of the superposition model for the redundant signals effect in bimodal detection. *Perception & Psychophysics*, *50*, 83–83.

Diederich, A., & Colonius, H. (2007a). Modeling spatial effects in visual-tactile reaction time. *Perception and Psychophysics*, *69*(1), 56–67.

Diederich, A., & Colonius, H. (2007b). Why two "distractors" are better then one: Modeling the effect on non-target auditory and tactile stimuli on visual saccadic reaction time. *Experimental Brain Research*, *179*, 43–54.

Diederich, A., & Colonius, H. (2008a). Crossmodal interaction in saccadic reaction time:

Separating multisensory from warning effects in the time window of integration model. *Experimental Brain Research*, *186*(1), 1–22.

Diederich, A., & Colonius, H. (2008b). When a high-intensity "distractor" is better then a low-intensity one: Modeling the effect of an auditory or tactile nontarget stimulus on visual saccadic reaction time. *Brain Research*, *1242*, 219–230.

Diederich, A., & Colonius, H. (2015). The time window of multisensory integration: Relating reaction times and judgments of temporal order. *Psychological Review*, *122*(2), 232–241.

Diederich, A., & Mallahi-Karai. (2018). Stochastic methods for modeling decision-making. In W. Batcheldeer, H. Colonius, & E. Dzhafarov (Eds.), *New handbook of mathematical psychology* (Vol. II, pp. 1–70). Cambridge: Cambridge University Press.

Diederich, A., & Oswald, P. (2016). Multi-stage sequential sampling models with finite or infinite time horizon and variable boundaries. *Journal of Mathematical Psychology*, *74*, 128–145.

Drugowitsch, J., DeAngelis, G., Klier, E. M., Angelaki, D., & Pouget, A. (2014). Optimal multisensory decision-making in a reaction-time task. *eLife*, e03005. (doi: 10.7554/eLife.03005)

Dzhafarov, E., & Colonius, H. (2006). Reconstructing distances among objects from their discriminability. *Psychometrika*, *71*(2), 365–386.

Dzhafarov, E., & Colonius, H. (2007). Dissimilarity cumulation theory and subjective metrics. *Journal of Mathematical Psychology*, *51*, 290–304.

Ernst, M., & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont donnés. *Annales de l'Université de Lyon*, *Section A, Séries 3*(14), 53–77.

Gondan, M. (2010). A permutation test for the race model inequality. *Behavior Research Methods*, *42*, 23–28.

Gondan, M., Dupont, D., & Blurton, S. (2020). Testing the race model in a difficult redundant signals task. *Journal of Mathematical Psychology*, *95*, https://doi.org/10.1016/j.jmp.2020.102323.

Gondan, M., Riehl, V., & Blurton, S. (2012). Showing that the race model inequality is not violated. *Behavior Research Methods*, *44*, 248–255.

Gondan, M., & Vorberg, D. (2021). Testing trisensory interactions. *Journal of Mathematical Psychology*, *101*, https://doi.org/10.1016/j.jmp.2021.102513.

Grant, K. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (l). *Journal of the Acoustical Society of America*, *112*(1), 30–33.

Grant, K., & Seitz, P. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America*, *104*(4), 2438–2450.

Green, D., & Swets, J. (1974). *Signal detection theory and psychophysics*. New York: Krieger.

Joe, H. (1997). *Multivariate models and dependence concepts* (No. 73). London: Chapman & Hall.

Jones, P. (2016). A tutorial on cue combination and signal detection theory: Using changes in sensitivity to evaluate how observers integrate sensory information. *Journal of Mathematical Psychology*, *73*, 117–139.

Lombardi, L., D'Allesandro, M., & Colonius, H. (2019). A new nonparametric test for the racemodel inequality. *Behavior Research Methods*, *51*, 2290–2301.

Lovelace, C., Stein, B., & Wallace, M. (2003). An irrelevant light enhances auditory detection in humans: A psychophysical analysis of multisensory integration in stimulus detection. *Cognitive Brain Research*, *17*, 447–453.

Macmillan, N., & Creelman, C. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mallahi-Karai, K., & Diederich, A. (2021). Decision with multiple alternatives: Geometric models in higher dimensions–the disk model. *Journal of Mathematical Psychology*, *100*(https://doi.org/10.1016/j.jmp.2020.102493).

Massaro, D., & Cohen, M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception. *Journal of the Acoustical Society of America*, *108*(2), 784–789.

Meredith, M., & Stein, B. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*, *221*, 389–391.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247–279.

Miller, R., Pluta, S., Stein, B., & Rowland, B. (2015). Relative unisensory strength and timing predict their multisensory product. *The Journal of Neuroscience*, *35*(13), 5213–5220.

Murray, M., & Wallace, M. (Eds.). (2012). *The neural bases of multisensory processes*. Boca Rato, FL: CRC Press.

Naumer, M., & Kayser, J. (Eds.). (2010). *Multisensory object perception in the primate brain*. New York: Springer-Verlag.

Otto, T., Dassy, B., & Mamassian, P. (2013). Principles of multisensory behavior. *The Journal of Neuroscience*, *33*, 7463–7474.

Raab, D. (1962). Statistical facilitation of simple reaction time. *Transactions of the New York Academy of Sciences*, *24*, 574–590.

Ratcliff, R. (1978). Theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ross, S. (1996). *Stochastic processes*, 2nd ed. New York: John Wiley & Sons.

Schwarz, W. (1989). A new model to explain the redundant-signals effect. *Perception & Psychophysics*, *46*(5), 490–500.

Schwarz, W. (1994). Diffusion, superposition, and the redundant-targets effect. *Journal of Mathematical Psychology*, *38*, 504–520.

Schwarz, W., & Miller, J. (2014). When less equals more: Probability summation without sensitivity improvement. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 2091–2100.

Smith, P. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408–463.

Stein, B., & Meredith, M. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stein, B., Stanford, T., Ramachandran, R., Perrault Jr, T., & Rowland, B. (2009). Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. *Experimental Brain Research*, *198*, 113–126.

Stein, B. E. (Ed.). (2012). *The new handbook of multisensory processes*. Cambridge, MA: MIT Press.

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P., Meredith, M., Perrault Jr, T., … Lewkowicz, D. (2010). Semantic confusion regarding the development of

multisensory integration: A practical solution. *European Journal of Neuroscience*, *31*, 1713–1720.

Stevenson, R., Ghose, D., Krueger Fister, J., Sarko, D., Altieri, N., Nidiffer, A., … Wallace, M. (2014). Identifying and quantifying multisensory integration: A tutorial review. *Brain Topography*, *27*(6), 707–730.

Todd, J. (1912). Reaction to multiple stimuli. *Archives of Psychology No. 25. Columbia Contributions to Philosophy and Psychology*, *XXI*(8). (New York: The Science Press.)

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, *28*(5), 656–668.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing*, *31*, 636–644.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*(4), 380–389.

Ulrich, R., Miller, J., & Schröter, H. (2007). Testing the race model inequality: An algorithm and computer programs. *Behavior Research Methods*, *39*(2), 291–302.

van de Rijt, L., Roye, A., Mylanus, E., van Opstal, A., & van Wanrooij, M. (2019, doi: 10.3389/fnhum.2019.00335). The principle of inverse effectiveness in audiovisual speech perception. *Frontiers in Human Neuroscience*, *13*(335).

van Opstal, A. (2016). The auditory system and human sound-localization behavior. In A. van Opstal (Ed.), *The auditory system and human sound-localization behavior* (pp. 361–392). San Diego, CA: Academic Press.

Welch, R., & Warren, D. (1986). Intersensory interactions. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 1–36). New York: John Wiley & Sons.

Wickens, T. (2002). *Elementary signal detection theory*. Oxford: Oxford University Press.

Yu, L., Cuppini, C., Xu, J., Rowland, B., & Stein, B. (2019). Cross-modal competition: The default computation for multisensory processing. *The Journal of Neuroscience*, *39*(8), 1374–1385.

3 Fechnerian Scaling: Dissimilarity Cumulation Theory

Ehtibar N. Dzhafarov and Hans Colonius

3.1 Introduction

3.1.1 What is it About?

In 1860 Gustav Theodor Fechner published the two-volume *Elemente der Psychophysik*. From this event one can date scientific psychology, firmly grounded in mathematics and experimental evidence. One of the main ideas introduced in Fechner's book is that of measuring subjective differences between stimuli **a** and **b** by means of summing (or integrating) just noticeable (or infinitesimal) differences in the interval of stimuli separating **a** and **b**. For Fechner, stimuli of a given kind are always represented by positive reals, so that the interval between them is well-defined.

We use the term "Fechnerian scaling" to designate any method of computing distances in a stimulus space by means of *cumulating* (summing, integrating) values of a *dissimilarity function* for pairs of "neighboring" stimuli. The term "dissimilarity cumulation" can be used as a synonym of "Fechnerian scaling" or else as designating an abstract mathematical theory of which Fechnerian scaling is the main application.

A *stimulus space* is a set of stimuli endowed with a structure imposed on this set by an observer's judgments. Thus, a set of all visible aperture colors such that for each pair of colors we have a number indicating how often they appear identical to an observer if presented side by side is an example of a stimulus space. Stimuli in a stimulus space are referred to as its *points*, and generally are denoted by boldface lowercase letters: $\mathbf{x}_k, \mathbf{a}, \mathbf{b}^{(\omega)}$, etc. The dissimilarity function is a generalization of the notion of a *metric*, mapping pairs of stimuli $(\mathbf{x}, \mathbf{y})$ into non-negative numbers $D(\mathbf{x}, \mathbf{y})$. On a very general level, with minimal assumptions about the structure of a stimulus space being considered, Fechnerian scaling is implemented by summing pairwise dissimilarities $D(\mathbf{x}_1, \mathbf{x}_2)$, $D(\mathbf{x}_2, \mathbf{x}_3)$, etc. along *finite chains* of points $\mathbf{a} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1} = \mathbf{b}$. The distance from **a** to **b** is then computed as the infimum of these cumulated values over the set of all such chains. Thus obtained distances from **a** to **b** and from **b** to **a** need not be the same, and to obtain a conventional, symmetric distance, in Fechnerian scaling one adds these distances together.

In more specialized stimulus spaces, finite chains can be replaced with continuous or even continuously differentiable *paths*. In the latter case the cumulation is replaced with integration along a path of a certain quantity, *submetric function*

$F(\mathbf{x}, \mathbf{u})$, that depends on the location $\mathbf{x}$ of a point and the velocity $\mathbf{u}$ with which it moves along the path. The submetric function is a measure of local discriminability of $\mathbf{x}$ from its "immediate" neighbors $\mathbf{x} + \mathbf{u}dx$, and it can be empirically estimated by means of one of Fechner's methods for measuring differential thresholds. The original methods are based on one's ability to compare stimuli in terms of "greater than" with respect to some property (brightness, loudness, extent, etc.). In more general situations, stimuli can be compared by a variety of methods based on one's ability to judge whether two stimuli are the same or different.

The structure defining a stimulus space on a set of stimuli is always imposed by an observer's judgments of the stimuli rather than by the way stimuli are measured as physical objects. In this sense, the structure of stimulus space is a psychological rather than a physical construct. For instance, a drawing of a human face has a complex physical description, but if, for example, faces are compared in terms of greater–less with respect to some property, such as "beauty," then, provided certain assumptions are satisfied, a set of all possible face drawings may form a unidimensional continuum mappable on an interval of reals. However, physical descriptions of the stimuli typically have some properties (e.g., order, closeness) suggestive of the respective properties of the judgments. For instance, if **a** and **b** have very similar physical descriptions, one can usually expect the results of their comparisons with any stimulus **c** to also be very similar – the consideration we use, for example, in constructing a differential–geometric version of Fechnerian scaling.

3.1.2 Unidimensional Fechnerian Scaling

Various aspects of Fechner's original theory are subject to competing interpretations because they are not presented in his writings with sufficient clarity. The following therefore is not a historical account. Rather, it is a modern theory that preserves the spirit of Fechner's idea of cumulation of small differences.

Let us assume that stimuli of a particular kind are represented (labeled, encoded) by values on an interval of positive real numbers $[t, u[$, where t is the absolute threshold value, and u is an appropriately defined upper threshold, or infinity. (Throughout this chapter, half-open or open intervals of reals will always be presented in the form $[t, u[$, $]t, u]$, $]t, u[$, using only square brackets.) The space structure on $[t, u[$ is defined by a *psychometric function* $\gamma(\mathbf{x}, \mathbf{y})$ that gives us the probability with which a stimulus $\mathbf{y}$ (represented by a value $y \in [t, u[$) is judged to be greater than stimulus $\mathbf{x}$ (represented by a value $x \in [t, u[$). In this special case, it is convenient to simply replace stimuli with their representations, and write x, y in place of $\mathbf{x}, \mathbf{y}$:

$$\gamma(x, y) = \Pr\left[y \text{ is judged to be greater than } x\right]. \tag{3.1}$$

We will make the simplifying assumption that

$$\gamma(y, x) = 1 - \gamma(x, y), \tag{3.2}$$

with the consequence

$$\gamma(x,x) = 1/2. \tag{3.3}$$

This will allow us to proceed in this special case without introducing the notions of observation areas and canonical transformations that are fundamental for the general theory.

Next, we will make a relatively innocuous assumption that $\gamma(x,y)$ is strictly increasing in y in the vicinity of $y = x$, and that it is continuously differentiable in y at $y = x$. That is, the derivative

$$F(x) = \left.\frac{\partial \gamma(x,y)}{\partial y}\right|_{y=x} \tag{3.4}$$

exists, is positive, and continuous in x. This is the slope of the psychometric function at its median, and the intuitive meaning of the differential $F(x)\,dx$ is that it is proportional to the dissimilarity between x and its "immediate" neighbor, $x+dx$. We can write this as

$$D(x, x+dx) = cF(x)\,dx,$$

where c is a positive constant specific to a given stimulus space. The intuition of cumulation of differences in this unidimensional setting is captured by the summation property

$$D(a,b) = D(a,c) + D(c,b),$$

for any $a \le c \le b$ in stimulus set $\mathfrak{S}$. It follows that

$$D(a,b) = c\int_a^b F(x)\,dx. \tag{3.5}$$

This quantity can be interpreted as the subjective distance between a and b for any $a \le b$ in $\mathfrak{S}$. We take the relations (3.4) and (3.5) for the core of the Fechnerian scaling in stimulus continua (presented here with simplifying assumptions).

3.1.3 Historical Digression: Fechner's Law

One can easily check that the *logarithmic law* advocated by Fechner,

$$D(t,x) = K\log\frac{x}{t}, x \ge t, \tag{3.6}$$

where K is a positive constant, corresponds to

$$F(x) = \frac{K}{x}, \tag{3.7}$$

which can be viewed as a differential form of the so-called *Weber's law*. Recall that t designates absolute threshold.

This is an example of the so-called *psychophysical law*, the relationship between a physical description of a stimulus x, and the value of $D(x,t)$, referred to as

the *magnitude of sensation*. In this chapter we attach little importance to this or other psychophysical laws. In view of the generalization of Fechnerian scaling to stimulus spaces with more complex descriptions than real numbers, such laws have limited scope of applicability.

Nevertheless, it is appropriate to take a historical detour and look at how Fechner's law was justified by Fechner himself, in the second volume of his landmark work *Elemente der Psychophysik*. The relationship (3.6) is referred to by Fechner as the *measurement formula (Massformel).* More generally, Fechner's law can be written as

$$D(a,b) = D(t,b) - D(t,a) = K \log \frac{b}{a}, b \geq a \geq t, \tag{3.8}$$

for two stimulus magnitudes a, b. Fechner calls this the *difference formula* (*Unterschiedsformel*).

In an addendum to his work *Zen Avesta*, Fechner describes how the idea of this law occurred to him on the morning of October 22, 1950 (this date is nowadays celebrated as the *Fechner Day*): he had an insight that an arithmetic progression of sensation magnitude should correspond to a geometric progression of stimulus magnitude. Fechner's insight on that day is all one needs to derive the law, as logarithm is the only function with non-chaotic behavior that can transform a geometric progression into an arithmetic one. The derivation of the law, however, had to wait 10 more years before it appeared in Volume 2 of the *Elemente der Psychophysik*, in two different forms (Chapters 16 and 17).

Unfortunately, the second volume has not been translated into English. As we learn from a letter written by E. G. Boring to S. Rosenzweig on February 23, 1968, "Just now I'm spending long hours working over translation into English of the second volume of the Fechner's *Elemente*, because put literally into English it is about as dull and confusing and sometimes uninterpretable as it always was in the German. Holt, Rinehart and Winston published the first volume and someday we will get this second half done, but we do not have much help after NIH stopped supporting translation. We have to get it done by little bits." It seems that Boring has not completed this work.

By a historical happenstance, one of Fechner's derivations of his law was criticized as mathematically incorrect, and the other simply forgotten. In addition, the law itself was criticized as empirically incorrect. However, by careful examination of the premises of Fechner's derivations the mathematical criticisms can be deflected, while empirical falsifications of the law often involve empirical procedures (e.g., direct estimation of sensation magnitudes) that go beyond those Fechner would consider legitimate. In a paper of rejoinders published in 1877, Fechner reacts to the criticisms known to him and makes a bold prediction for the future: "The tower of Babel was never finished because the workers could not reach an understanding on how they should build it; my psychophysical edifice will stand because the workers will never agree on how to tear it down."

The difficulty in understanding Fechner's derivations of his logarithmic law is that he uses the term "Weber's law" in the meaning that is logically independent of the empirical law established by Ernst Heinrich Weber (which Fechner, to add to the confusion, also calls "Weber's law"). According to Weber's law, if x and $x+\Delta x$ are separated by a *just-noticeable difference*, then

$$\frac{\Delta x}{x} = c^*, \tag{3.9}$$

where c^* is a constant with respect to x (but generally depends on the stimulus continuum used). In Fechner's mathematical derivations, however, the term "Weber's law" stands for the following statement, essentially a form of his October 1850 insight:

> *the subjective dissimilarity $D(t,b) - D(t,a)$ between stimuli with physical magnitudes a and b (provided $t \leq a \leq b$) is determined by the ratio of these magnitudes, b/a.*

We propose calling this statement the "W-principle" to disentangle it from Weber's law. The only relationship between the W-principle and Weber's law can be established through the so-called "Fechner's postulate," according to which all just-noticeable differences Δx (within a given continuum) are subjectively equal:

$$D(x, x+\Delta x) = c. \tag{3.10}$$

Any two of the three statements, Fechner's postulate, Weber's law (in its usual meaning), and the W-principle, imply the third.

In Chapter 17 of the *Elemente*, Fechner derives his law by using a novel (for his time) method of *functional equations*. He presents the W-principle as

$$\psi(b) - \psi(a) = F\left(\frac{b}{a}\right),$$

where $\psi(x)$ denotes $D(t,x)$, and observes that this implies

$$F\left(\frac{c}{b}\right) + F\left(\frac{b}{a}\right) = F\left(\frac{c}{a}\right),$$

for any $t \leq a \leq b \leq c$. This in turn means that

$$F(x) + F(y) = F(xy),$$

for any $x, y \geq 1$. Fechner recognizes in this the functional equation introduced only 40 years earlier by Augustin-Louis Cauchy, who showed that its only continuous solution is

$$F(x) = K \log x, x \geq 1.$$

It is now known (Aczél, 1987) that continuity can be replaced with many other regularity assumptions, including monotonicity and non-negativity, and that it is

sufficient to assume that these properties hold only in an arbitrarily small vicinity of 1 (i.e., for very weak stimuli only). It follows that

$$\psi(b) - \psi(a) = K \log \frac{b}{a}, b \geq a \geq t,$$

which is Fechner's *Unterschiedsformel*.

In Chapter 16 of the *Elemente*, Fechner derives the same relationship in a different way. He presents the functional equation as

$$\psi(b) - \psi(a) = G\left(\frac{b-a}{a}\right),$$

and by assuming that G is differentiable at zero gets the differential equation

$$\psi'(x)\, dx = G'(0)\, \frac{dx}{x},$$

whose integration once again leads to Fechner's logarithmic formula.

The novelty of the method of functional equations in the mid-nineteenth century is probably responsible for the fact that the Chapter 17 derivation was universally overlooked by Fechner's contemporaries (and then, as it seems, forgotten altogether). The derivation in Chapter 16, through differential equations, was, by contrast, common in Fechner's time, which may be the reason Fechner placed it first. This derivation has been criticized as mathematically or logically flawed by Fechner's contemporaries and modern authors alike. The common interpretation has been that it is based on Fechner's postulate

$$\psi(x + \Delta x) - \psi(x) = c.$$

He is thought to have combined this with Weber's law

$$\frac{\Delta x}{x} = c^*,$$

to arrive at

$$\psi(x + \Delta x) - \psi(x) = \frac{c}{c^*}\frac{\Delta x}{x}.$$

Finally, Fechner is thought to have invoked an "expediency principle" (*Hülfsprinzip*) to illegitimately replace the finite differences with differentials:

$$d\psi = \frac{c}{c^*}\frac{dx}{x}.$$

The integration of this equation with the boundary condition $\psi(x_0) = 0$ yields

$$\psi(x) = \frac{c}{c^*} \log \frac{x}{x_0}.$$

It has been pointed out that this derivation is internally contradictory because it implies

$$\psi(x + \Delta x) - \psi(x) = \frac{c}{c^*} \log \frac{x + \Delta x}{x} = \frac{c}{c^*} \log\left(1 + c^*\right),$$

which is not the same as the postulated

$$\psi(x+\Delta x)-\psi(x)=c.$$

Boring's characterization of Fechner's book as "dull and confusing and sometimes uninterpretable" being true, it is not easy to refute this criticism. However, it is clear that Fechner uses neither the Fechner postulate nor Weber's law in deriving his law, although he accepts the truth of both. As explained above, he makes use of the W-principle (which he calls "Weber's law"). It follows from his derivation that if Weber's law holds in addition to the W-principle, then

$$\psi(x+\Delta x)-\psi(x)=K\log\left(1+c^*\right)=c,$$

which is indeed a constant (Fechner's postulate proved as a theorem). As Fechner points out in a book of rejoinders, if the Weber fraction c^* is sufficiently small, the constant K *approximately* equals c/c^*, as in the criticized formula. The "expediency principle" which Fechner's critics especially disparage seems to be nothing more than an inept and verbose explanation of the elementary fact (used in the Chapter 16 derivation) that if a function $f(x)$ is differentiable at zero, then $df(x)$ is proportional to dx.

3.1.4 Observation Areas and Canonical Transformation

The elementary but fundamental fact is that if an observer is asked to compare two stimuli, $\mathbf{x}$ and $\mathbf{y}$, they must differ in some respect that allows the observer to identify them as two distinct stimuli. For instance, in the pair written as $(\mathbf{x},\mathbf{y})$, the first argument, $\mathbf{x}$, may denote the stimulus presented chronologically first, followed by $\mathbf{y}$. Or $\mathbf{x}$ may always be presented above or to the left of $\mathbf{y}$. In perceptual pairwise comparisons, the stimuli must differ in their spatial and/or temporal locations, but the defining properties of $\mathbf{x}$ and $\mathbf{y}$ in the pair $(\mathbf{x},\mathbf{y})$ may vary. Thus, two line segments to be compared in length may be presented in varying pairs of distinct spatial locations, but one of the line segments may always be vertical (and written first in the pair, $\mathbf{x}$) and the other horizontal (written second, $\mathbf{y}$).

Formally, this means that a stimulus space involves two stimulus sets rather than one. Denoting them $\mathfrak{S}_1^{\star\star}$ (for $\mathbf{x}$-stimuli) and $\mathfrak{S}_2^{\star\star}$ (for $\mathbf{y}$-stimuli), we call them the first and the second *observation areas,* respectively. The space structure is imposed on the Cartesian product of these observation areas by a function

$$\phi^{\star\star}:\mathfrak{S}_1^{\star\star}\times\mathfrak{S}_2^{\star\star}\to R, \tag{3.11}$$

where R may be a set of possible responses, or possible probabilities of a particular response.

We say that two stimuli $\mathbf{x},\mathbf{x}'\in\mathfrak{S}_1^{\star\star}$ are *psychologically equal* if

$$\phi^{\star\star}(\mathbf{x},\mathbf{y})=\phi^{\star\star}\left(\mathbf{x}',\mathbf{y}\right)$$

for any $\mathbf{y}\in\mathfrak{S}_2^{\star\star}$. Similarly, $\mathbf{y},\mathbf{y}'\in\mathfrak{S}_2^{\star\star}$ are psychologically equal if

$$\phi^{\star\star}(\mathbf{x},\mathbf{y})=\phi^{\star\star}\left(\mathbf{x},\mathbf{y}'\right),$$

for any $\mathbf{x} \in \mathfrak{S}_1^{\star\star}$. One can always relabel the elements of the observation areas by assigning identical labels to all psychologically equal stimuli. For instance, all metameric colors may be encoded by the same RGB coordinates irrespective of their spectral composition. Objects of different color but of the same weight will normally be labeled identically in a task involving hefting and deciding which of two objects is heavier.

Let us denote by $\mathfrak{S}_1^{\star}$ and $\mathfrak{S}_2^{\star}$ the observation areas in which psychologically equal stimuli are equal. The function $\phi^{\star\star}$ is then redefined into

$$\phi^{\star} : \mathfrak{S}_1^{\star} \times \mathfrak{S}_2^{\star} \to R. \tag{3.12}$$

We will illustrate this transformation by a toy example. Let the original function be

ϕ^*	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$
$\mathbf{x}_1$	0.7	0.6	0.3	0.4	0.4	0.4	0.4
$\mathbf{x}_2$	0.5	0.3	0.4	0.2	0.2	0.2	0.2
$\mathbf{x}_3$	0.5	0.3	0.4	0.2	0.2	0.2	0.2
$\mathbf{x}_4$	0.2	0.1	0.5	0.3	0.3	0.3	0.3
$\mathbf{x}_5$	0.2	0.1	0.5	0.3	0.3	0.3	0.3
$\mathbf{x}_6$	0.1	0.3	0.8	0.6	0.6	0.6	0.6
$\mathbf{x}_7$	0.1	0.3	0.8	0.6	0.6	0.6	0.6

The first observation area, $\mathfrak{S}_1^{\star\star}$, comprises stimuli $\{\mathbf{x}_1, \ldots, \mathbf{x}_7\}$ (e.g., weights placed on one's left palm), the second observation area, $\mathfrak{S}_2^{\star\star}$, comprises stimuli $\{\mathbf{y}_1, \ldots, \mathbf{y}_7\}$ (weights placed on one's right palm), and the entries in the matrix above are values of $\phi^*(\mathbf{x}, \mathbf{y})$, an arbitrary function mapping $(\mathbf{x}, \mathbf{y})$-pairs into real numbers (say, the probabilities of deciding that the two weights differ in heaviness). If two rows (or columns) of the matrix are identical, then the two corresponding $\mathbf{x}$-stimuli (respectively, $\mathbf{y}$-stimuli) are psychologically equal, and can be labeled identically. Thus, the stimuli in $\{\mathbf{x_2}, \mathbf{x_3}\}$, in $\{\mathbf{x_4}, \mathbf{x_5}\}$, in $\{\mathbf{x_6}, \mathbf{x_7}\}$, and in $\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}$ are psychologically equal and they can be replaced by a single symbol, respectively. The redefined spaces $\mathfrak{S}_1^{\star}$ and $\mathfrak{S}_1^{\star}$ are then as follows:

$$\begin{array}{cccccccc} \mathfrak{S}_1^{\star\star}: & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 & \mathbf{x}_5 & \mathbf{x}_6 & \mathbf{x}_7 \\ & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow \\ \mathfrak{S}_1^{\star}: & \mathbf{x}_a & \mathbf{x}_b & \mathbf{x}_b & \mathbf{x}_c & \mathbf{x}_c & \mathbf{x}_d & \mathbf{x}_d \end{array}, \quad \begin{array}{cccccccc} \mathfrak{S}_2^{\star\star}: & \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 & \mathbf{y}_4 & \mathbf{y}_5 & \mathbf{y}_6 & \mathbf{y}_7 \\ & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow \\ \mathfrak{S}_2^{\star}: & \mathbf{y}_a & \mathbf{y}_b & \mathbf{y}_c & \mathbf{y}_d & \mathbf{y}_d & \mathbf{y}_d & \mathbf{y}_d \end{array}$$

and the function ϕ^* transforms into $\phi^{\star}$ accordingly:

$\phi^{\star}$	$\mathbf{y}_a$	$\mathbf{y}_b$	$\mathbf{y}_c$	$\mathbf{y}_d$
$\mathbf{x}_a$	0.7	0.6	0.3	0.4
$\mathbf{x}_b$	0.5	0.3	0.4	0.2
$\mathbf{x}_c$	0.2	0.1	0.5	0.3
$\mathbf{x}_d$	0.1	0.3	0.8	0.6

As another example, consider the function $\gamma(x,y)$ of the previous section, and assume that

$$\mathfrak{S}_1^{\star\star} = [t_1, u_1[, \mathfrak{S}_2^{\star\star} = [t_2, u_2[.$$

Assume that $\gamma(x,y)$ is strictly increasing in y and strictly decreasing in x. Then $\gamma(x,y) = \gamma(x,y')$ implies $y = y'$ and $\gamma(x,y) = \gamma(x',y)$ implies $x = x'$, so that in this case

$$\mathfrak{S}_1^{\star\star} = \mathfrak{S}_1^{\star}, \mathfrak{S}_2^{\star\star} = \mathfrak{S}_2^{\star}.$$

Staying with this example, $\gamma(x,y) = 1/2$ defines here the binary relation "is matched by": $x \in \mathfrak{S}_1^{\star}$ is matched by $y \in \mathfrak{S}_2^{\star}$ if and only if $\gamma(x,y) = 1/2$. The relation "$y \in \mathfrak{S}_2^{\star}$ is matched by $x \in \mathfrak{S}_1^{\star}$" is defined by the same condition, $\gamma(x,y) = 1/2$. The traditional psychophysical designation of this relation is that y is the *point of subjective equality* (PSE) for x (and then x is the PSE for y). The assumptions (3.2) and (3.3) made in the previous section do not hold generally. In particular, the psychometric function γ, as a rule, has a nonzero *constant error* (i.e., $\gamma(x,y) = 1/2$ does not imply $x = y$; see Figure 3.1).

With the monotonicity assumptions about γ made above, if we also assume that the range of the function $y \mapsto \gamma(x,y)$ for every x includes the value $1/2$, and that the same is true for the range of the function $x \mapsto \gamma(x,y)$ for every y, then we have the following properties of the PSE relation (see Figure 3.2):

1. the PSE for every $x \in \mathfrak{S}_1^{\star}$ exists and is unique;
2. the PSE for every $y \in \mathfrak{S}_2^{\star}$ exists and is unique;
3. $y \in \mathfrak{S}_2^{\star}$ is a PSE for $x \in \mathfrak{S}_1^{\star}$ if and only if $x \in \mathfrak{S}_1^{\star}$ is a PSE for $y \in \mathfrak{S}_2^{\star}$.

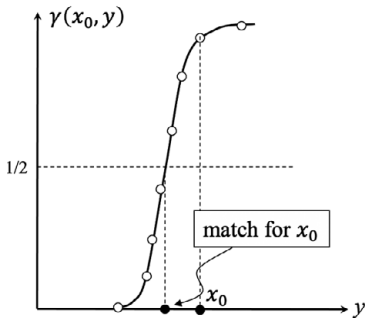


Figure 3.1 *A "greater–less" psychometric function $y \mapsto \gamma(x,y)$ defined on an interval of real numbers. The function shows, for a fixed value of $x = x_0$, the probability with which y is judged to be greater than x_0 with respect to some designated property. The median value of y, one at which $\gamma(x_0,y) = \frac{1}{2}$, is taken to be a match, or point of subjective equality (PSE) for x_0, and the difference between x_0 and its PSE defines constant error. (Note that showing $\gamma(x,y)$ at a fixed value of x does not mean that the value of x was fixed procedurally in an experiment. The graph is simply a cross-section of $\gamma(x,y)$ at $x = x_0$.)*

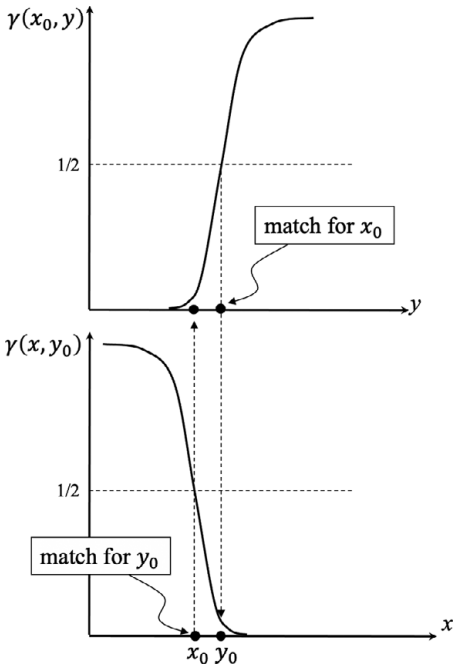


Figure 3.2 *An illustration, for the psychometric function $\gamma\,(x,y)$, of the symmetry of the relation "to be a PSE for." The upper panel shows the function $y \mapsto \gamma\,(x,y)$ at $x = x_0$, and y_0 denotes the PSE for x_0. The lower panel shows the function $x \mapsto \gamma\,(x,y)$ at $y = y_0$, and x_0 then has to be the PSE for y_0. Conversely, if x_0 denotes the PSE for y_0 in the lower panel, then y_0 has to be the PSE for x_0 in the upper panel. This follows from the fact that in both cases the PSE is defined by $\gamma\,(x,y) = \frac{1}{2}$, and the assumption that both $x \mapsto \gamma\,(x,y)$ and $y \mapsto \gamma\,(x,y)$ are monotone functions whose range includes the value $\gamma = \frac{1}{2}$.*

We will assume that these properties generalize to any function $\phi^\star$ in (3.12). In other words, we assume that $\phi^\star$ is associated with a *bijective function* $\mathbf{h} : \mathfrak{S}_1^\star \longrightarrow \mathfrak{S}_2^\star$ such that for all $\mathbf{x} \in \mathfrak{S}_1^\star$ and $\mathbf{y} \in \mathfrak{S}_2^\star$,

(P1) $\mathbf{y}$ is a PSE for $\mathbf{x}$ if and only if $\mathbf{y} = \mathbf{h}\,(\mathbf{x})$;
(P2) $\mathbf{x}$ is a PSE for $\mathbf{y}$ if and only if $\mathbf{x} = \mathbf{h}^{-1}\,(\mathbf{y})$.

This makes the relation of "being a PSE of" or "being matched by" symmetric. As a result, one can always apply to the observation areas a *canonical transformation*

$$\mathbf{f} : \mathfrak{S}_1^\star \longrightarrow \mathfrak{S}, \mathbf{g} : \mathfrak{S}_2^\star \longrightarrow \mathfrak{S},$$

with $\mathbf{f}$ and $\mathbf{g}$ arbitrary except for

$$\mathbf{h} = \mathbf{g}^{-1} \circ \mathbf{f}.$$

A canonical transformation redefines the function $\phi^\star$ into

$$\phi : \mathfrak{S} \times \mathfrak{S} \longrightarrow R,$$

such that, for any ordered pair $(\mathbf{x},\mathbf{y})$, one of the elements is a PSE for the other element if and only if $\mathbf{x}=\mathbf{y}$. We say that the stimulus space and the space-forming function ϕ here are in a *canonical form*.

Let us use the toy example above for an illustration. We assume that the PSE for any $\mathbf{x}$ is defined here as $\mathbf{y}$ at which $\mathbf{y}\mapsto\phi^{\star}(\mathbf{x},\mathbf{y})$ reaches its minimum; and the PSE for any $\mathbf{y}$ is defined as $\mathbf{x}$ at which $\mathbf{x}\mapsto\phi^{\star}(\mathbf{x},\mathbf{y})$ reaches its minimum. The inspection of the matrix for $\phi^{\star}$ shows that the PSEs are well defined for both $\mathbf{x}$-stimuli and $\mathbf{y}$-stimuli:

$\phi^{\star}$	$\mathbf{y}_a$	$\mathbf{y}_b$	$\mathbf{y}_c$	$\mathbf{y}_d$
$\mathbf{x}_a$	0.7	0.6	**0.3**	0.4
$\mathbf{x}_b$	0.5	0.3	0.4	**0.2**
$\mathbf{x}_c$	0.2	**0.1**	0.5	0.3
$\mathbf{x}_d$	**0.1**	0.3	0.8	0.6

We also see that in each row the minimal value (shown boxed) is also minimal in its column. That is, $\mathbf{y}$ is a PSE for $\mathbf{x}$ if and only if $\mathbf{x}$ is the PSE for $\mathbf{y}$. The graph of the bijective $\mathbf{h}$-function in the formulations of the properties P1 and P2 is given by the pairs

$$\{(\mathbf{x}_a,\mathbf{y}_c),(\mathbf{x}_b,\mathbf{y}_d),(\mathbf{x}_c,\mathbf{y}_b),(\mathbf{x}_d,\mathbf{y}_a)\}.$$

Simple relabeling then allows us to have all PSE pairs on the main diagonal. Both $\mathfrak{S}_1^{\star}$ and $\mathfrak{S}_1^{\star}$ can be mapped into one and the same set $\mathfrak{S}$, for example,

$$\begin{array}{cccccc} \mathfrak{S}_1^{\star}: & \mathbf{x}_a & \mathbf{x}_b & \mathbf{x}_c & \mathbf{x}_d & \\ & \Downarrow & \Downarrow & \Downarrow & \Downarrow & , \\ \mathfrak{S}: & \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{d} & \end{array} \qquad \begin{array}{cccccc} \mathfrak{S}_2^{\star}: & \mathbf{y}_c & \mathbf{y}_d & \mathbf{y}_b & \mathbf{y}_a & \\ & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \\ \mathfrak{S}: & \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{d} & \end{array}$$

and $\phi^{\star}$ transforms into ϕ accordingly:

ϕ	**a**	**b**	**c**	**d**
a	**0.3**	0.4	0.6	0.7
b	0.4	**0.2**	0.3	0.5
c	0.5	0.3	**1**	0.2
d	0.8	0.6	0.3	**0.1**

To apply canonical transformation to our second example, the psychometric function $\gamma(x,y)$, let us assume that $\gamma(x,y)=1/2$ holds if and only if $y=h(x)$ for some homeomorphic mapping h (i.e., such that both h and h^{-1} are continuous). Since $\mathfrak{S}_1^{*}=\mathfrak{S}_1^{\star}=[t_1,u_1[$ and $\mathfrak{S}_2^{*}=\mathfrak{S}_2^{\star}=[t_2,u_2[$, $\mathfrak{S}$ can always be chosen in the form $[t,u[$, by choosing any two homeomorphisms

$$f:[t_1,u_1[\to[t,u[, g:[t_2,u_2[\to[t,u[,$$

such that $g^{-1}\circ f\equiv h$. Note, however, that this only ensures compliance with (3.3), but not with (3.2).

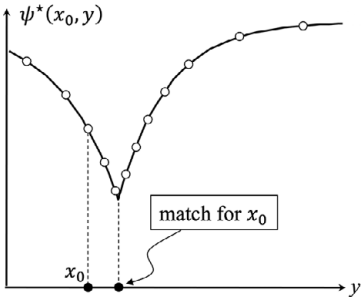


Figure 3.3 *A "same–different" psychometric function $y \mapsto \psi^\star(x, y)$ defined on an interval of real numbers. The function shows, for a fixed value of $x = x_0$, the probability with which y is judged to be different from x_0 (generically or with respect to a designated property). The value of y at which $\psi^\star(x_0, y)$ reaches its minimum is taken to be a match, or point of subjective equality (PSE) for x_0.*

3.1.5 Same–Different Judgments

The greater–less comparisons are possible only with respect to a designated characteristic, such as loudness or beauty. It is clear, however, that no such characteristic can reflect all relevant aspects of the stimuli being compared. Moreover, it is not certain that the characteristic's values are always comparable in terms of greater–less, given a sufficiently rich stimulus set. Thus, it may not be clear to an observer which of two given faces is more beautiful, and even loudness may not be semantically unidimensional if the sounds are complex. The same–different comparisons have a greater scope of applicability, and do not have to make use of designated characteristics. The role of the stimulus-space-defining function ϕ^* of the previous section in this case is played by

$$\psi^*(\mathbf{x}, \mathbf{y}) = \Pr\left[\mathbf{y} \text{ is judged to be different from } \mathbf{x}\right], \tag{3.13}$$

with $\mathbf{x} \in \mathfrak{S}_1^{\star\star}$ and $\mathbf{y} \in \mathfrak{S}_2^{\star\star}$. To be different here means to differ in any respect other than the conspicuous difference between the two observation areas. Thus, if $\mathbf{x}$ is a visual stimulus always presented to the left of $\mathbf{y}$, this difference in spatial locations does not enter in the judgments of whether $\mathbf{x}$ and $\mathbf{y}$ are different or the same. Of course, it is also possible to ask whether the two stimuli differ in a particular respect, such as color or shape.

The reduction of $\left(\psi^*, \mathfrak{S}_1^{\star\star}, \mathfrak{S}_2^{\star\star}\right)$ to $\left(\psi^\star, \mathfrak{S}_1^\star, \mathfrak{S}_2^\star\right)$, in which psychologically equal stimuli are equal, is effected by assigning an identical label to any $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}_1^{\star\star}$ such that

$$\psi^{\star\star}(\mathbf{x}, \mathbf{y}) = \psi^{\star\star}\left(\mathbf{x}', \mathbf{y}\right)$$

for all $\mathbf{y} \in \mathfrak{S}_2^{\star\star}$, and similarly for the second observation area.

The PSE relation for the function $\psi^\star$ is defined as follows (see Figure 3.4): $\mathbf{y} \in \mathfrak{S}_2^\star$ is a PSE for $\mathbf{x} \in \mathfrak{S}_1^\star$ if

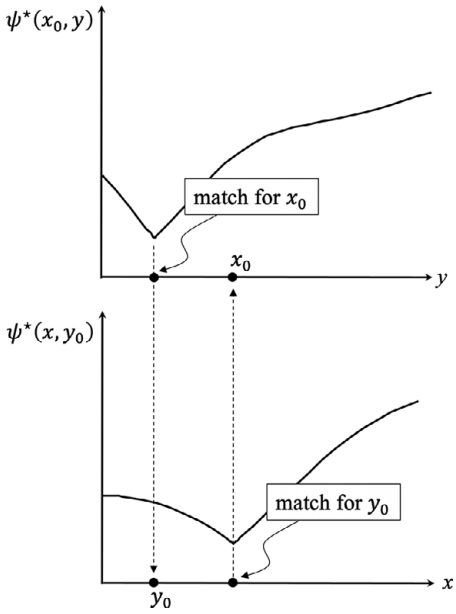


Figure 3.4 *An illustration, for the psychometric function $\psi^\star(x,y)$ in Figure 3.3, of the symmetry of the relation "to be a PSE for." The upper panel shows the function $y \mapsto \psi^\star(x,y)$ at $x = x_0$, and y_0 denotes the PSE for x_0. The lower panel shows the function $x \mapsto \psi^\star(x,y)$ at $y = y_0$, and x_0 is shown to be PSE for y_0. Conversely, if x_0 denotes the PSE for y_0 in the lower panel, then y_0 is shown to be the PSE for x_0 in the upper panel. Unlike in the case of the "greater/less" psychometric function (Figure 3.2), here the symmetry of the PSE relation is an assumption rather than a consequence of other properties of the function $\psi^\star$.*

$$\psi^\star(\mathbf{x},\mathbf{y}) < \psi^\star\left(\mathbf{x},\mathbf{y}'\right) \text{ for all } \mathbf{y}' \neq \mathbf{y}.$$

Analogously, $\mathbf{x} \in \mathfrak{S}_1^\star$ is a PSE for $\mathbf{y} \in \mathfrak{S}_2^\star$ if

$$\psi^\star(\mathbf{x},\mathbf{y}) < \psi^\star\left(\mathbf{x}',\mathbf{y}\right) \text{ for all } \mathbf{x}' \neq \mathbf{x}.$$

In accordance with the previous section, we assume the existence of a bijection $\mathbf{h}: \mathfrak{S}_1^\star \longrightarrow \mathfrak{S}_2^\star$ such that

$$\begin{array}{c} \psi^\star(\mathbf{x},\mathbf{h}(\mathbf{x})) < \psi^\star(\mathbf{x},\mathbf{y}) \text{ for all } \mathbf{y} \neq \mathbf{h}(\mathbf{x}), \\ \psi^\star\left(\mathbf{h}^{-1}(\mathbf{y}),\mathbf{y}\right) < \psi^\star(\mathbf{x},\mathbf{y}) \text{ for all } \mathbf{x} \neq \mathbf{h}^{-1}(\mathbf{y}). \end{array} \tag{3.14}$$

That is, we assume that the PSEs in the space $\left(\psi^\star, \mathfrak{S}_1^\star, \mathfrak{S}_2^\star\right)$ exist, are unique, and that $\mathbf{y}$ is the PSE for $\mathbf{x}$ if and only if $\mathbf{x}$ is the PSE for $\mathbf{y}$. We refer to this property as the *law of regular minimality*. In this chapter it should be taken as part of the definition of the functions we are dealing with rather than an empirical claim.

Now, any canonical transformation, as described above, yields a probability function

$$\psi : \mathfrak{S} \times \mathfrak{S} \longrightarrow [0,1], \tag{3.15}$$

such that, for any $\mathbf{a}, \mathbf{x}, \mathbf{y} \in \mathfrak{S}$, if $\mathbf{x} \neq \mathbf{a}$ and $\mathbf{y} \neq \mathbf{a}$, then

$$\psi(\mathbf{a}, \mathbf{a}) < \begin{cases} \psi(\mathbf{x}, \mathbf{a}) \\ \psi(\mathbf{a}, \mathbf{y}) \end{cases}. \tag{3.16}$$

We will assume in the following that the discrimination probability function ψ is presented in this canonical form. This by no means implies that $\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{y}, \mathbf{x})$, the order of the arguments continues to matter. We will continue to consider the two arguments in $\psi(\mathbf{x}, \mathbf{y})$ as belonging to the first and second observation areas, respectively.

3.2 Notation Conventions

We now introduce notation conventions for the rest of this chapter. They in part codify and in part modify the notation used in the introductory section.

Let us agree that from now on real-valued functions of one or several points of a stimulus set will be indicated by strings without parentheses: $\psi\mathbf{ab}$ in place of $\psi(\mathbf{a}, \mathbf{b})$, $D\mathbf{abc}$ in place of $D(\mathbf{a}, \mathbf{b}, \mathbf{c})$, etc. Boldface lowercase letters denoting stimuli are merely labels, with no implied operations between them, so this notation is unambiguous. (In Section 3.8, lowercase boldface letters are also used to denote direction vectors, in which case the string convention is not used.) If a stimulus is represented by a real number we may conveniently confuse the two, and write, for example, $\gamma(x, y)$ instead of the more rigorous $\gamma\mathbf{xy}$ with $\mathbf{x}, \mathbf{y}$ represented by (or having values) x, y.

A finite sequence (or *chain*) $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of points in a stimulus set will be presented as a string $\mathbf{x}_1 \ldots \mathbf{x}_n$. If a chain of stimuli is to be referred to without indicating its elements, then it is indicated by uppercase boldface letters. Thus $\mathbf{X}$ may stand for $\mathbf{abc}$, $\mathbf{Y}$ stand for $\mathbf{y}_1 \ldots \mathbf{y}_n$, etc. If $\mathbf{X} = \mathbf{x}_1 \ldots \mathbf{x}_k$ and $\mathbf{Y} = \mathbf{y}_1 \ldots \mathbf{y}_l$ are two chains, then

$$\begin{gathered} \mathbf{XY} = \mathbf{x}_1 \ldots \mathbf{x}_k \mathbf{y}_1 \ldots \mathbf{y}_l, \\ \mathbf{aXb} = \mathbf{a}\mathbf{x}_1 \ldots \mathbf{x}_k \mathbf{b}, \\ \mathbf{aXbYa} = \mathbf{a}\mathbf{x}_1 \ldots \mathbf{x}_k \mathbf{b}\mathbf{y}_1 \ldots \mathbf{y}_l \mathbf{a}, \\ \text{etc.} \end{gathered}$$

The number of elements in a chain $\mathbf{X}$ is its cardinality $|\mathbf{X}|$. Infinite sequences $\{x_1, \ldots, x_n, \ldots\}$, $\{\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots\}$, $\{\mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots\}$, etc., are almost always indicated by their generic elements: numerical sequence $\{x_n\}$, stimulus sequence $\{\mathbf{x}_n\}$, sequence of chains $\{\mathbf{X}_n\}$, etc. Convergence of a sequence, such as $\mathbf{x}_n \to \mathbf{x}$, is understood as conditioned on $n \to \infty$. In a sequence of chains, the cardinality $|\mathbf{X}_n|$ is generally changing.

As mentioned earlier, we indicate intervals of reals (closed, open, and half-open) by square brackets: $[a, b]$, $[a, b[$, $]a, b]$, and $]a, b[$. Round-bracketed pairs of numbers of stimuli, (a, b) or $(\mathbf{a}, \mathbf{b})$, always indicate an ordered pair.

Sets of stimuli are denoted by Gothic letters, $\mathfrak{S}$, $\mathfrak{S}_1^{\star\star}$, $\mathfrak{s}$, etc. For sets of chains and paths in stimulus spaces we use script letters, $\mathcal{C}, \mathcal{P}_a^b$, etc. For other types of

sets we use blackboard and sans serif fonts on an ad hoc basis. The set of reals is denoted, as usual, $\mathbb{R}$.

3.3 Basics of Fechnerian Scaling

Using our new notation, and considering an at least two-element stimulus space $\mathfrak{S}$ in a canonical form, we have, for any distinct points $\mathbf{x}$ and $\mathbf{y}$ in $\mathfrak{S}$,

$$\begin{aligned} \Psi^{(1)}\mathbf{xy} &= \psi\mathbf{xy} - \psi\mathbf{xx} > 0, \\ \Psi^{(2)}\mathbf{xy} &= \psi\mathbf{yx} - \psi\mathbf{xx} > 0. \end{aligned} \tag{3.17}$$

We call the quantities $\Psi^{(1)}\mathbf{xy}$ and $\Psi^{(2)}\mathbf{xy}$ *psychometric increments* of the first and second kind, respectively. Both can be interpreted as ways of quantifying the intuition of a dissimilarity of $\mathbf{y}$ from $\mathbf{x}$. The order "from–to" is important here, as $\Psi^{(i)}\mathbf{yx} \neq \Psi^{(i)}\mathbf{yx}$ $(i = 1, 2)$.

In Fechnerian scaling we use the psychometric increments to compute subjective distances in the spirit of Fechner's idea of cumulation of small dissimilarities. We will see that this cumulation can assume different forms, depending on the properties of a stimulus space. However, the general construction, applicable to all spaces, is as follows.

3.3.1 Step 1

First, we assume that both $\Psi^{(1)}$ or $\Psi^{(2)}$ are *dissimilarity functions*, in accordance with the following definition (to be explained and elaborated later on).

Definition 3.1. We say that $D : \mathfrak{S} \times \mathfrak{S} \to \mathbb{R}$ is a dissimilarity function if it has the following properties:

$\mathcal{D}1$(*positivity*) $D\mathbf{ab} > 0$ for any distinct $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$;

$\mathcal{D}2$ (*zero property*) $D\mathbf{aa} = 0$ for any $\mathbf{a} \in \mathfrak{S}$;

$\mathcal{D}3$ (*uniform continuity*) for any $\varepsilon > 0$ one can find a $\delta > 0$ such that, for any $\mathbf{a}, \mathbf{b}, \mathbf{a}', \mathbf{b}' \in \mathfrak{S}$,

$$\text{if } D\mathbf{aa}' < \delta \text{ and } D\mathbf{bb}' < \delta, \text{ then } \left|D\mathbf{a}'\mathbf{b}' - D\mathbf{ab}\right| < \varepsilon;$$

$\mathcal{D}4$ (*chain property*) for any $\varepsilon > 0$ one can find a $\delta > 0$ such that for any chain $\mathbf{aXb}$,

$$\text{if } D\mathbf{aXb} < \delta, \text{ then } D\mathbf{ab} < \varepsilon.$$

For the chain property, we need to define $D\mathbf{aXb}$.

Definition 3.2. Given a chain $\mathbf{X} = \mathbf{x}_1 \dots \mathbf{x}_k$ in $\mathfrak{S}$, its *D-length* (or just *length* once D is specified) is defined as

$$D\mathbf{X} = \begin{cases} D\mathbf{x}_1\mathbf{x}_2 + \dots + D\mathbf{x}_{k-1}\mathbf{x}_k & \text{if } |\mathbf{X}| > 1 \\ 0 & \text{if } |\mathbf{X}| \leq 1 \end{cases}.$$

Then, for a given pair of points $\mathbf{a}, \mathbf{b}$, the length of $\mathbf{aXb}$ is

$$D\mathbf{aXb} = \begin{cases} D\mathbf{a}\mathbf{x}_1 + D\mathbf{X} + D\mathbf{x}_k\mathbf{b} & \text{if } |\mathbf{X}| > 0 \\ D\mathbf{ab} & \text{if } |\mathbf{X}| = 0 \end{cases}.$$

3.3.2 Step 2

Next, we consider the set $\mathcal{C}$ of all (finite) chains in $\mathfrak{S}$,

$$\mathcal{C} = \bigcup_{k=0}^{\infty} \mathfrak{S}^k,$$

and define

$$G\mathbf{ab} = \inf_{\mathbf{X} \in \mathcal{C}} D\mathbf{aXb}. \tag{3.18}$$

We will see below that the function $G : \mathfrak{S} \times \mathfrak{S} \to \mathbb{R}$ is a *quasimetric dissimilarity*, in accordance with the following definition.

Definition 3.3. Function $M : \mathfrak{S} \times \mathfrak{S} \to \mathbb{R}$ is a quasimetric dissimilarity function if it has the following properties:

$\mathcal{QM}1$ (*positivity*) $M\mathbf{ab} > 0$ for any distinct $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$;

$\mathcal{QM}2$ (*zero property*) $M\mathbf{aa} = 0$ for any $\mathbf{a} \in \mathfrak{S}$;

$\mathcal{QM}3$ (*triangle inequality*) $M\mathbf{ab} + M\mathbf{bc} \geq M\mathbf{ac}$ for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathfrak{S}$;

$\mathcal{QM}4$ (*symmetry in the small*) for any $\varepsilon > 0$ one can find a $\delta > 0$ such that $M\mathbf{ab} < \delta$ implies $M\mathbf{ba} < \varepsilon$, for any $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$.

To relate quasimetric dissimilarity to two familiar terms, a function satisfying $\mathcal{QM}1$–$\mathcal{QM}3$ is called a *quasimetric*, and a quasimetric is called a *metric* if it satisfies the property

$\mathcal{M}4$ (symmetry) $M\mathbf{ab} = M\mathbf{ba}$, for any $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$.

Quasimetric dissimilarity therefore can be viewed as a concept intermediate between quasimetric and metric. More importantly, however, a quasimetric dissimilarity (hence also a metric), as shown below, is a special form of dissimilarity, whereas quasimetric generally is not (see Figure 3.5).

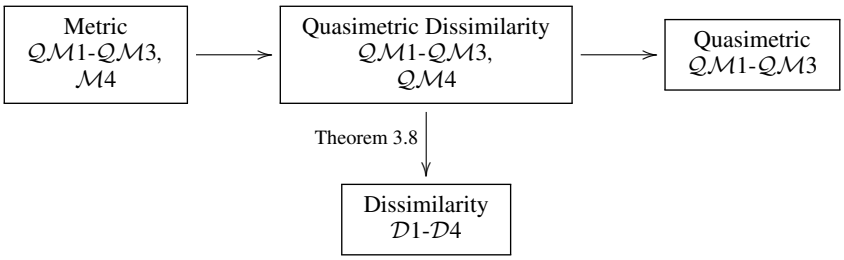


Figure 3.5 *Interrelations between metric-like concepts. Arrows between the boxes stand for "is a special case of."*

3.3.3 Step 3

The quasimetric dissimilarities

$$G^{(1)}\mathbf{ab} = \inf_{\mathbf{X}\in\mathcal{C}} \Psi^{(1)}\mathbf{aXb}$$

and

$$G^{(2)}\mathbf{ab} = \inf_{\mathbf{X}\in\mathcal{C}} \Psi^{(2)}\mathbf{aXb}$$

are generally different. However, we will see below that

$$G^{(1)}\mathbf{ab} + G^{(1)}\mathbf{ba} = G^{(2)}\mathbf{ab} + G^{(2)}\mathbf{ba}, \tag{3.19}$$

and this quantity is clearly a metric. We will denote it $\overleftrightarrow{G}\mathbf{ab}$, and interpret it as the *Fechnerian distance* between **a** and **b** in the canonical stimulus space $\mathfrak{S}$. The double-arrow in $\overleftrightarrow{G}$ is suggestive of the following way of presenting this quantity:

$$\overleftrightarrow{G}\mathbf{ab} = \inf_{\mathbf{X},\mathbf{Y}\in\mathcal{C}} \Psi^{(1)}\mathbf{aXbYa} = \inf_{\mathbf{X},\mathbf{Y}\in\mathcal{C}} \Psi^{(2)}\mathbf{aXbYa}, \tag{3.20}$$

the **aXbYa** (equivalently, **bYaXb**) being a closed chain containing the points **a** and **b**.

3.3.4 Subsequent Development

The function $\overleftrightarrow{G}$ is, in a sense, the ultimate goal of Fechnerian scaling. However, the metric structure of a space is part of its geometry, and this is what a full theory of Fechnerian scaling deals with. In discrete spaces, consisting of isolated points, the general definition of $\overleftrightarrow{G}$ provides the algorithm for computing it. In more structured spaces, however, the Fechnerian metric may be computed in specialized ways. Rather than considering all possible chains, in some spaces one integrates infinitesimal dissimilarities along continuous paths and seeks the shortest paths. In still more structured spaces this leads to a generalized form of Finsler geometry, where computations of distances are based on indicatrices or submetric functions.

The psychometric increments $\Psi^{(1)}$ and $\Psi^{(2)}$ are at the foundation of Fechnerian scaling. In this chapter they are defined through the psychometric function ψ in (3.13), which is usually associated with the same–different version of the *method of constant stimuli*. In this method, same–different judgments are recorded for repeatedly presented multiple pairs of stimuli, as indicated, for example, by the open circles in Figure 3.3. However, virtually any pairwise comparison procedure can be, in principle, used to define analogs of $\Psi^{(1)}$ and $\Psi^{(2)}$. For instance, if the observer judges pairs of stimuli in terms of "greater–less" with respect to some property, the psychometric function γ of Figure 3.1 (assuming it is in a canonical form) can be converted into a ψ-like function by putting

$$\psi\mathbf{xy} = \begin{cases} \gamma\mathbf{xy} & \text{if } \gamma\mathbf{xy} \geq \frac{1}{2} \\ 1-\gamma\mathbf{xy} & \text{if } \gamma\mathbf{xy} < \frac{1}{2} \end{cases}.$$

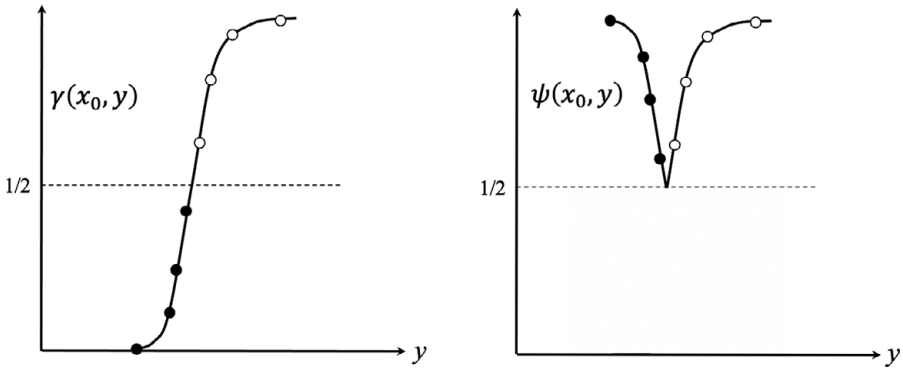


Figure 3.6 *An illustration of how a "greater–less" discrimination probability function (on the left) can be redefined into a "same–different"-type discrimination probability function.*

This is illustrated in Figure 3.6 for the case $\mathfrak{S}$ is an interval of real numbers. The psychometric increments then are defined as

$$\Psi^{(1)}\mathbf{xy} = \left|\gamma\mathbf{xy} - \frac{1}{2}\right|, \Psi^{(2)}\mathbf{xy} = \left|\gamma\mathbf{yx} - \frac{1}{2}\right|.$$

Some experimental procedures may yield dissimilarity values $D\mathbf{ab}$ "directly." Thus, in one of the procedures of *multidimensional scaling* (MDS), observers are presented pairs of stimuli and asked to numerically estimate "how different they are." Then, for every pair of stimuli $\mathbf{a}, \mathbf{b}$, some measure of central tendency of these numerical estimates can be hypothesized to be an efficient estimator of a dissimilarity function

$$\Psi^{(1)}\mathbf{ab} = \Psi^{(2)}\mathbf{ba} = D\mathbf{ab}.$$

If one can establish that $D\mathbf{aa} = 0$ for all stimuli and that $D\mathbf{ab} > 0$ for distinct $\mathbf{a}, \mathbf{b}$, then the stimulus space is in a canonical form, and the hypothesis that D is a dissimilarity function cannot be falsified on any finite set of data. However, given sufficient amount of data, one can usually falsify the hypothesis that $D\mathbf{ab}$ is a quasimetric, by establishing that $D\mathbf{ab}$ violates the triangle inequality. In such situations, MDS seeks a monotone transformation $g \circ D$ that would yield a quasimetric. Dissimilarity cumulation offers an alternative approach, to use D to compute by (3.18) a quasimetric dissimilarity G, and then symmetrize it by (3.20). We will return to this situation in Section 3.9.

3.4 Dissimilarity Function

The properties $\mathcal{D}3$ and $\mathcal{D}4$ of Definition 3.1 are more conveniently presented in terms of convergence of sequences. Let us introduce convergence in a stimulus space.

Definition 3.4. Given two sequences of points in $\mathfrak{S}$, $\{\mathbf{a}_n\}$ and $\{\mathbf{b}_n\}$, we say that $\mathbf{a}_n$ and $\mathbf{b}_n$ *converge to each other, and write this* $\mathbf{a}_n \leftrightarrow \mathbf{b}_n$, if $D\mathbf{a}_n\mathbf{b}_n \to 0$. In the special case $\mathbf{b}_n \equiv \mathbf{b}$, we say that $\mathbf{a}_n$ *converges to* $\mathbf{b}$, and write $\mathbf{a}_n \to \mathbf{b}$.

The property $\mathcal{D}3$ (uniform continuity) can then be presented as follows:

$$\text{if } \mathbf{a}_n \leftrightarrow \mathbf{a}'_n \text{ and } \mathbf{b}_n \leftrightarrow \mathbf{b}'_n, \text{ then } D\mathbf{a}'_n\mathbf{b}'_n - D\mathbf{a}_n\mathbf{b}_n \to 0.$$

In other words, D is a uniformly continuous function (Figure 3.7).

It is clear that $\mathbf{a}_n \leftrightarrow \mathbf{a}_n$ is true for any sequence $\{\mathbf{a}_n\}$ (because $D\mathbf{a}_n\mathbf{a}_n = 0$). Assuming that $\mathbf{a}_n \leftrightarrow \mathbf{b}_n$, we can use $\mathcal{D}3$ to observe that

$$\mathbf{a}_n \leftrightarrow \mathbf{b}_n \text{ and } \mathbf{a}_n \leftrightarrow \mathbf{a}_n \Longrightarrow D\mathbf{a}_n\mathbf{a}_n - D\mathbf{b}_n\mathbf{a}_n \to 0 \Longleftrightarrow D\mathbf{b}_n\mathbf{a}_n \to 0.$$

But $D\mathbf{b}_n\mathbf{a}_n \to 0$ means $\mathbf{b}_n \leftrightarrow \mathbf{a}_n$, and we obtain the following proposition.

Theorem 3.5 (*symmetry in the small*) *For any* $\{\mathbf{a}_n\}$, $\{\mathbf{b}_n\}$,

$$\mathbf{a}_n \leftrightarrow \mathbf{b}_n \text{ iff } \mathbf{b}_n \leftrightarrow \mathbf{a}_n.$$

This justifies the terminology (convergence to each other) and notation in the definition of $\mathbf{a}_n \leftrightarrow \mathbf{b}_n$.

Property $\mathcal{D}4$ (chain property) can be presented as follows: for any sequences $\{\mathbf{a}_n\}$, $\{\mathbf{b}_n\}$ in $\mathfrak{S}$ and $\{\mathbf{X}_n\}$ in $\mathcal{C}$ (the set of chains),

$$\text{if } D\mathbf{a}_n\mathbf{X}_n\mathbf{b}_n \to 0, \text{ then } \mathbf{a}_n \leftrightarrow \mathbf{b}_n. \tag{3.21}$$

Figure 3.8 provides an illustration.

The properties $\mathcal{D}1$–$\mathcal{D}4$ are logically independent: none of them is a consequence of the remaining three. This is proved by constructing examples, for each of these properties, that violate this property while conforming to the others. For example, to prove the independence of $\mathcal{D}4$, consider $\mathfrak{S} = \mathbb{R}$, and let $D\mathbf{xy} = (x-y)^2$ (where x, y are the numerical values representing $\mathbf{x}, \mathbf{y}$, respectively). The function D clearly satisfies $\mathcal{D}1$–$\mathcal{D}3$. However, for any points $\mathbf{a}, \mathbf{b}$, if the elements of a chain $\mathbf{X}_n$ subdivide $[a, b]$ into n equal parts, then

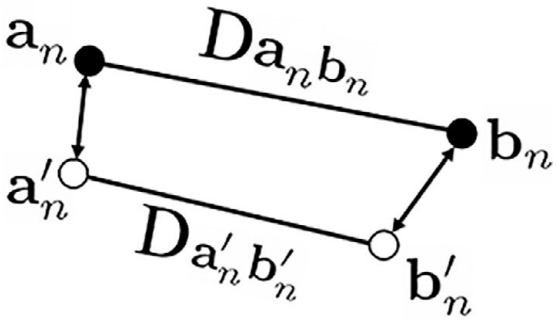


Figure 3.7 *Illustration of the uniform continuity of D. The dissimilarities* $D\mathbf{a}_n\mathbf{b}_n$ *and* $D\mathbf{a}'_n\mathbf{b}'_n$ *converge to each other as* $\mathbf{a}_n$ *with* $\mathbf{a}'_n$ *converge to each other and* $\mathbf{b}_n$ *with* $\mathbf{b}'_n$ *converge to each other.*

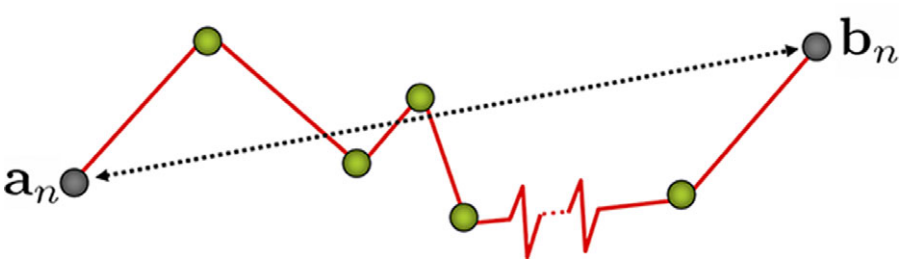


Figure 3.8 *Illustration of the chain property of D. If the overall length of the chains* $\mathbf{X}_n$ *connecting* $\mathbf{b}_n$ *to* $\mathbf{a}_n$ *tends to zero, then* $\mathbf{a}_n$ *and* $\mathbf{b}_n$ *converge to each other. This property is nontrivial only if* $|\mathbf{X}_n|$*, the number of elements in the chains, tends to infinity. If it is bounded,* $\mathbf{a}_n \leftrightarrow \mathbf{b}_n$ *is a consequence of the transitivity of the* $\leftrightarrow$ *relation (not discussed in the text, but easily established).*

$$D\mathbf{a}\mathbf{X}_n\mathbf{b} = n\left(\frac{b-a}{n}\right)^2 \to 0,$$

while the value of $D\mathbf{ab}$ remains equal to $(b-a)^2$.

3.5 Quasimetric Dissimilarity

We begin by establishing an important fact: the function G defined by (3.18) and the dissimilarity D are *equivalent in the small*.

Theorem 3.6. *For any* $\{\mathbf{a}_n\}, \{\mathbf{b}_n\}$,

$$\mathbf{a}_n \leftrightarrow \mathbf{b}_n \text{ iff } G\mathbf{a}_n\mathbf{b}_n \to 0.$$

To prove this, we first observe that $G\mathbf{ab} \geq 0$, as the infimum of non-negative $D\mathbf{a}\mathbf{X}\mathbf{b}$. If $D\mathbf{a}_n\mathbf{b}_n \to 0$, we have

$$0 \leq G\mathbf{a}_n\mathbf{b}_n = \inf_{\mathbf{X}\in\mathcal{C}} D\mathbf{a}_n\mathbf{X}\mathbf{b}_n \leq D\mathbf{a}_n\mathbf{b}_n \to 0,$$

and this implies $G\mathbf{a}_n\mathbf{b}_n \to 0$. Conversely, $\inf_{\mathbf{X}\in\mathcal{C}} D\mathbf{a}_n\mathbf{X}\mathbf{b}_n \to 0$ means that for some sequence of chains $\{\mathbf{X}_n\}$, $D\mathbf{a}_n\mathbf{X}_n\mathbf{b}_n \to 0$. By the chain property then, $D\mathbf{a}_n\mathbf{b}_n \to 0$.

Let us now see if G satisfies the properties defining a quasimetric dissimilarity, $\mathcal{QM}1$–$\mathcal{QM}4$. We immediately see that it satisfies the triangle inequality ($\mathcal{QM}3$):

$$G\mathbf{ab} \leq G\mathbf{ac} + G\mathbf{cb},$$

for any $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathfrak{S}$. Indeed

$$G\mathbf{ac} + G\mathbf{cb} = \inf_{\mathbf{X}\in\mathcal{C}} D\mathbf{a}\mathbf{X}\mathbf{c} + \inf_{\mathbf{Y}\in\mathcal{C}} D\mathbf{c}\mathbf{Y}\mathbf{b} = \inf_{\mathbf{X},\mathbf{Y}\in\mathcal{C}} D\mathbf{a}\mathbf{X}\mathbf{c}\mathbf{Y}\mathbf{b},$$

and the set of all possible $\mathbf{aXb}$ contains the set of all possible $\mathbf{aXcYb}$ chains. It is also easy to see that the function G is symmetric in the small ($\mathcal{QM}4$). Written in convergence terms, the property is

$$\text{if } G\mathbf{a}_n\mathbf{b}_n \to 0 \text{ then } G\mathbf{b}_n\mathbf{a}_n \to 0.$$

It is proved by observing that, by the previous theorem, if $G\mathbf{a}_n\mathbf{b}_n \to 0$ then $\mathbf{a}_n \leftrightarrow \mathbf{b}_n$, and then $G\mathbf{b}_n\mathbf{a}_n \to 0$. Because we know that $G\mathbf{ab}$ is non-negative, the properties $\mathcal{QM}1$ and $\mathcal{QM}2$ follow from

$$G\mathbf{ab} = \inf_{\mathbf{X}\in\mathcal{C}} D\mathbf{aXb} = 0 \Longrightarrow D\mathbf{aX}_n\mathbf{b} \to 0,$$

for some sequence of chains $\{\mathbf{X}_n\}$. But this means, by the chain property, $D\mathbf{ab} = 0$, which is true if and only if $\mathbf{a} = \mathbf{b}$. We have established therefore

Theorem 3.7. *The function G is a quasimetric dissimilarity.*

It is instructive to see why, as mentioned earlier and as its name suggests, any quasimetric dissimilarity, and G in particular, is a dissimilarity function. Let M satisfy the properties $\mathcal{QM}1$–$\mathcal{QM}4$. Then $\mathcal{D}1$ and $\mathcal{D}2$ are satisfied trivially. The property $\mathcal{D}3$ (uniform continuity) follows from the fact that, by the triangle inequality,

$$\begin{cases} M\mathbf{aa}' + M\mathbf{b}'\mathbf{b} & \geq M\mathbf{ab} - M\mathbf{a}'\mathbf{b}', \\ M\mathbf{a}'\mathbf{a} + M\mathbf{bb}' & \geq M\mathbf{a}'\mathbf{b}' - M\mathbf{ab}. \end{cases}$$

By the symmetry in the small property,

$$\begin{aligned} M\mathbf{a}_n\mathbf{a}'_n \to 0 &\Longleftrightarrow M\mathbf{a}'_n\mathbf{a}_n \to 0, \\ M\mathbf{b}'_n\mathbf{b}_n \to 0 &\Longleftrightarrow M\mathbf{b}_n\mathbf{b}'_n \to 0, \end{aligned}$$

so these convergences imply

$$\left|M\mathbf{a}_n\mathbf{b}_n - M\mathbf{a}'_n\mathbf{b}'_n\right| \to 0.$$

The chain property, $\mathcal{D}4$, follows from $M\mathbf{aXb} \geq M\mathbf{ab}$, by the triangle inequality. We have established therefore

Theorem 3.8. *Any quasimetric dissimilarity (hence also any metric) is a dissimilarity function.*

Let us now return to the definition of $G^{(1)}$, $G^{(2)}$, and $\overleftrightarrow{G}$. We need to establish (3.20), from which (3.19) follows. Given a chain $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_k$, let us define the opposite chain $\mathbf{X}^\dagger$ as $\mathbf{x}_k\mathbf{x}_{k-1}\ldots\mathbf{x}_1$. By straightforward algebra:

$$\Psi^{(1)}\mathbf{X} = \sum_{i=1}^{k-1}\Psi^{(1)}\mathbf{x}_i\mathbf{x}_{i+1} = \sum_{i=1}^{k-1}\left(\psi\mathbf{x}_i\mathbf{x}_{i+1} - \psi\mathbf{x}_i\mathbf{x}_i\right),$$

$$\Psi^{(2)}\mathbf{X}^\dagger = \sum_{i=1}^{k-1}\Psi^{(2)}\mathbf{x}_{i+1}\mathbf{x}_i = \sum_{i=1}^{k-1}\left(\psi\mathbf{x}_i\mathbf{x}_{i+1} - \psi\mathbf{x}_{i+1}\mathbf{x}_{i+1}\right).$$

It follows that

$$\Psi^{(1)}\mathbf{X} - \Psi^{(2)}\mathbf{X}^\dagger = \psi\mathbf{x}_k\mathbf{x}_k - \psi\mathbf{x}_1\mathbf{x}_1.$$

In particular, if the chain is closed, $\mathbf{x}_k = \mathbf{x}_1$, we have

$$\Psi^{(1)}\mathbf{X} = \Psi^{(2)}\mathbf{X}^\dagger.$$

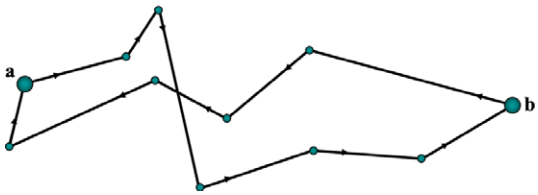


Figure 3.9 *For any closed chain* $\mathbf{X}$ *containing points* $\mathbf{a}, \mathbf{b}$*, the value of* $\Psi^{(1)}\mathbf{X}$ *is the same as the value of* $\Psi^{(2)}\mathbf{X}^{\dagger}$*, the same chain traversed in the opposite direction.*

That is, the $\Psi^{(1)}$-length of a closed chain equals the $\Psi^{(2)}$-length of the same chain traversed in the opposite direction (see Figure 3.9). Applying this to a chain $\mathbf{aXbYa}$,

$$\Psi^{(1)}\mathbf{aXbYa} = \Psi^{(2)}\mathbf{aY}^{\dagger}\mathbf{bX}^{\dagger}\mathbf{a},$$

whence

$$\inf_{\mathbf{X},\mathbf{Y}\in\mathcal{C}} \Psi^{(1)}\mathbf{aXbYa} = \inf_{\mathbf{Y}^{\dagger},\mathbf{X}^{\dagger}\in\mathcal{C}} \Psi^{(2)}\mathbf{aY}^{\dagger}\mathbf{bX}^{\dagger}\mathbf{a}.$$

Clearly, the set of all possible pairs of chains $(\mathbf{X}, \mathbf{Y})$ is the same as the set of all pairs $\left(\mathbf{Y}^{\dagger}, \mathbf{X}^{\dagger}\right)$, and by simple renaming,

$$\inf_{\mathbf{X},\mathbf{Y}\in\mathcal{C}} \Psi^{(1)}\mathbf{aXbYa} = \inf_{\mathbf{X},\mathbf{Y}\in\mathcal{C}} \Psi^{(2)}\mathbf{aXbYa}.$$

This proves.

Theorem 3.9. *For any* $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$*,*

$$G^{(1)}\mathbf{ab} + G^{(1)}\mathbf{ba} = G^{(2)}\mathbf{ab} + G^{(2)}\mathbf{ba} = \overleftrightarrow{G}\,\mathbf{ab}.$$

The function $\overleftrightarrow{G}$ *is a metric.*

The last statement is an immediate corollary of Theorem 3.7.

One can think of other ways of combining quasimetric dissimilarities $G^{(1)}\mathbf{ab}$ and $G^{(1)}\mathbf{ba}$ into a metric, such as

$$\max\left(G^{(1)}\mathbf{ab}, G^{(1)}\mathbf{ba}\right), \sqrt{G^{(1)}\mathbf{ab} + G^{(1)}\mathbf{ba}}, \text{etc.}$$

Denoting a combination like this $f\left(G^{(1)}\mathbf{ab}, G^{(1)}\mathbf{ba}\right)$, the natural requirements are that

(i) it should equal $f\left(G^{(2)}\mathbf{ab}, G^{(2)}\mathbf{ba}\right)$, and
(ii) $f(x,x) \propto x$.

The latter requirement ensures that if $G^{(1)}\mathbf{ab}$ always equals $G^{(1)}\mathbf{ba}$ (i.e., it is already a metric), then $f\left(G^{(1)}\mathbf{ab}, G^{(1)}\mathbf{ab}\right)$ is just a multiple of $G^{(1)}\mathbf{ab}$. Clearly, function $\overleftrightarrow{G}$ satisfies these requirements. In fact, up to a scaling coefficient, it is the only such function.

Theorem 3.10. *Function $f(x,y)$ satisfies **(i)** and **(ii)** above for all stimulus spaces if and only if $f(x,y) = k(x+y)$.*

For proof, consider a canonical space $(\psi, \mathfrak{S})$ with $\mathfrak{S} = \{\mathbf{a}, \mathbf{b}\}$. It is easy to see that for any $s, z \in]0,1]$ one can find probabilities $\psi\mathbf{aa}$, $\psi\mathbf{ab}$, $\psi\mathbf{ba}$, $\psi\mathbf{bb}$ satisfying

$$\begin{aligned} G_1\mathbf{ab} &= \psi\mathbf{ab} - \psi\mathbf{aa} = s, \\ G_1\mathbf{ba} &= \psi\mathbf{ba} - \psi\mathbf{bb} = s, \\ G_2\mathbf{ab} &= \psi\mathbf{ba} - \psi\mathbf{aa} = 2s - z, \\ G_2\mathbf{ba} &= \psi\mathbf{ab} - \psi\mathbf{bb} = z. \end{aligned}$$

Then the requirement **(i)** means that

$$f(s,s) = f(2s - z, z)$$

should hold for all $s, z \in]0,1]$. That is, $f(2s-z,z)$ depends on s only, and we have

$$f(x,y) = g(x+y).$$

Putting $x = y = \frac{u}{2}$, it follows from the requirement **(ii)** that

$$g(x+y) = g(u) = ku,$$

for some $k > 0$. So, our definition of $\overleftrightarrow{G}$ is not arbitrary, except for choosing $k = 1$.

3.6 Dissimilarity Cumulation in Discrete Spaces

3.6.1 Direct Computation of Distances

A discrete stimulus space $(\mathfrak{S}, D)$ consists of isolated points, that is, for every $\mathbf{x} \in \mathfrak{S}$,

$$\inf_{\mathbf{y} \in \mathfrak{S}, \mathbf{y} \neq \mathbf{x}} D\mathbf{xy} > 0. \tag{3.22}$$

Although genuinely discrete and even finite stimulus spaces exist (e.g., the Morse codes of letters and digits studied for their confusability), this special case is important not so much in its own right as because any set of empirical data forms a discrete (in fact, finite) space. This means, for example, that even if an observer is asked to compare colors or sounds, the data will form a finite set of pairs associated with some estimate of discriminability. If the data are sufficiently representative, the results of applying Fechnerian scaling of discrete spaces to them should provide a good approximation to the theoretical Fechnerian scaling using dissimilarity cumulation along continuous or smooth paths, as described later in this chapter.

As mentioned earlier, in discrete spaces the general definition of a Fechnerian distance directly determines the algorithm of computing them: one tries all possible chains leading from one point to another (with some obvious heuristics shrinking this set), and finds their infimum or, in special cases, minimum. This is illustrated in Figure 3.10.

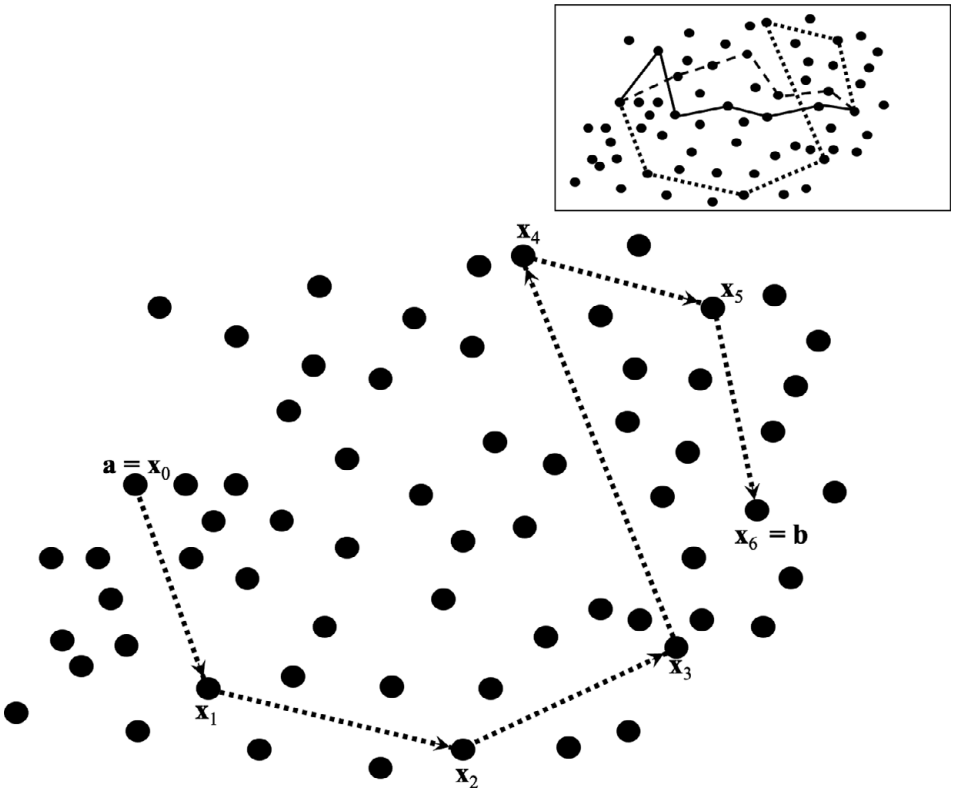


Figure 3.10 *Dissimilarity cumulation in discrete spaces. One considers all possible chains connecting a point* **a** *to a point* **b** *and seeks the infimum of their D-lengths. In a finite space this infimum is the smallest among the D-lengths, and it may be attained by more than one chain.*

Let us return to the toy example presented in Section 3.1.4, and assume that the function ϕ there is in fact the discrimination probability function ψ. The canonical space $(\mathfrak{S} = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}, \psi)$ is represented by the matrix that we reproduce here for convenience:

ψ	**a**	**b**	**c**	**d**
a	0.3	0.4	0.6	0.7
b	0.4	0.2	0.3	0.5
c	0.5	0.3	0.1	0.2
d	0.8	0.6	0.3	0.1

We know that all computations can be performed with either $\Psi^{(1)}$ or $\Psi^{(2)}$; the final result will be the same. Let us therefore compute $\Psi^{(1)}\mathbf{xy}$ by subtracting from each entry $\psi\mathbf{xy}$ the diagonal value in the same row, $\psi\mathbf{xx}$ (because the row labels are representing the stimuli in the first observation area). The result is

$\Psi^{(1)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.3	0.4
b	0.2	0	0.1	0.3
c	0.4	0.2	0	0.1
d	0.7	0.5	0.2	0

(3.23)

Let us consider next all chains leading from **a** to **d**, and from **d** to **a**. We obviously need not consider chains with loops in them (such as **abcacd**, containing loops **cac** and **abca**):

from **a** to **d**	$\Psi^{(1)}$-length
ad	0.4
abd	$0.1+0.3$
acd	$0.3+0.1$
abcd	$0.1+0.1+0.1$
acbd	$0.3+0.2+0.3$

,

from **d** to **a**	$\Psi^{(1)}$-length
da	0.7
dba	$0.5+0.2$
dca	$0.2+0.4$
dcba	$0.2+0.2+0.2$
dbca	$0.5+0.1+0.4$

The shortest chains here are **abcd** and either of **dca** and **dcba**, their $\Psi^{(1)}$-lengths being, respectively,

$$G^{(1)}\mathbf{ad} = 0.3, G^{(1)}\mathbf{da} = 0.6.$$

Thence

$$\overleftrightarrow{G}\,\mathbf{ad} = 0.3 + 0.6 = 0.9.$$

Repeating this procedure for each other pair of stimuli, we obtain the following complete set of $G^{(1)}$-distances,

$G^{(1)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.2	0.3
b	0.2	0	0.1	0.2
c	0.4	0.2	0	0.1
d	0.6	0.4	0.2	0

(3.24)

and, by symmetrization, the complete set of Fechnerian distances

$\overleftrightarrow{G}$	**a**	**b**	**c**	**d**
a	0	0.3	0.6	0.9
b	0.3	0	0.3	0.6
c	0.6	0.3	0	0.3
d	0.9	0.6	0.3	0

(3.25)

The shortest chains are not generally unique, as we have seen in our toy example. However, their infimum for any given pair of points (in the case of finite sets, minimum) is always determined uniquely. (Note that it is only a numerical accident that all $\overleftrightarrow{G}$ in our example are below 1, there is no general upper bound for $\overleftrightarrow{G}$ computed from probability values.)

Recall that a label in the canonical stimulus space, say, **a**, is a representation of two different stimuli in the two observation areas. If one goes back to the original stimulus spaces, the Fechnerian distance 0.6 between points **b** and **d** in the canonical space $\mathfrak{S}$ is in fact both

(i) the distance between either of the stimuli $\mathbf{x}_2, \mathbf{x}_3$ and either of the stimuli $\mathbf{x}_6, \mathbf{x}_7$ in the stimulus space $\mathfrak{S}_1^*$ (first observation area); and

(ii) the distance between any of the stimuli $\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7$ and the stimulus $\mathbf{y}_1$ in the stimulus space $\mathfrak{S}_2^*$ (second observation area).

Indeed, any of the stimuli $\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7$ and either of $\mathbf{x}_2, \mathbf{x}_3$ are each other's PSEs, mapped into **b** in the canonical representation. Similarly, either of the stimuli $\mathbf{x}_6, \mathbf{x}_7$ and $\mathbf{y}_1$ are each other's PSEs, mapped into **d**.

Let us emphasize that Fechnerian distances are always defined *within* observation areas rather than across them. This is the reason Fechnerian distance $\overleftrightarrow{G}$ is a true metric, with the symmetry property. Within a single observation area the order of two stimuli has no operational meaning, so $\overleftrightarrow{G}\,\mathbf{xy}$ cannot be different from $\overleftrightarrow{G}\,\mathbf{yx}$. The situation is different when we consider a discrimination probability function ψ or a dissimilarity function D (e.g., $\Psi^{(1)}$ or $\Psi^{(2)}$). In $\psi\mathbf{xy}$ and $D\mathbf{xy}$ the first and second stimuli belong to, respectively, the first and second observation areas, making them meaningfully asymmetric.

The quasimetric dissimilarity G (e.g., $G^{(1)}$or $G^{(2)}$) from which $\overleftrightarrow{G}$ is computed, strictly speaking, is not interpretable before it is symmetrized. $G\mathbf{xy}$ is merely a component of $\overleftrightarrow{G}\,\mathbf{xy}$, the other component being $G\mathbf{yx}$. However, in the rest of this paper we are focusing on G rather than $\overleftrightarrow{G}$ because the computation of G from D is the nontrivial part of Fechnerian scaling, leaving one only the trivial step of adding $G\mathbf{yx}$ to $G\mathbf{xy}$.

3.6.2 Recursive Corrections for Violations of the Triangle Inequality

The procedure described in this section is not the only way to compute G from D. Another way, known as the Floyd–Warshall algorithm, is based on the following logic. If one considers in $\mathfrak{S}$ all possible ordered triples **xyz** with pairwise distinct elements, and finds out that all of them satisfy the triangle inequality

$$D\mathbf{xz} \leq D\mathbf{xy} + D\mathbf{yz},$$

then D simply coincides with G. If therefore, in the general case, one could "correct" all ordered triples **xyz** for violations of the triangle inequality, one would transform D into G. The following is how this can be done for any finite stimulus space (a generalization to be discussed in Section 3.9.3).

Let $\mathfrak{S}$ contains k points, and let $\mathfrak{S}_3$ denote the set of $t = k(k-1)(k-2)$ ordered triples of pairwise distinct points of $\mathfrak{S}$. We will call the elements of $\mathfrak{S}_3$ *triangles*. For $n = 0, 1, \ldots$, let $\mathbf{T}^{(n)}$ denote a sequence of the t triangles in $\mathfrak{S}_3$ (in an arbitrary order, as its choice will be shown to be immaterial for the end result).

For each n, we index the triangles in $\mathbf{T}^{(n)}$ by double indices $(n,1)$, $(n,2)$, ..., (n,t), and we order all such pairs lexicographically: the successor $(n,i)'$ of (n,i) is $(n,i+1)$ if $i<t$ and $(n,t)'=(n+1,1)$. So the triangle indexed $(n,i)'$ is in $\mathbf{T}^{(n)}$, while the triangle indexed $(n,t)'$ is the first one in $\mathbf{T}^{(n+1)}$.

Definition 3.11. Given a finite space $(\mathfrak{S},D)$ and the triangle sequences $\mathbf{T}^{(0)},\mathbf{T}^{(1)},\ldots$, the dissimilarity function $M^{(n,i)}$ for $n=0,1,\ldots$ and $i=1,2,\ldots,t$ is defined by induction as follows.

(i) $M^{(0,i)}\equiv D$ for $i=1,2,\ldots,t$.

(ii) Let $M^{(n,i)}$ be defined for some $(n,i)\geq(0,t)$, and let $\mathbf{abc}$ be the triangle indexed by $(n,i)'$. Then $M^{(n,i)'}\mathbf{xy}=M^{(n,i)}\mathbf{xy}$ for all $\mathbf{x},\mathbf{y}\in\mathfrak{S}$ except, possibly, for $M^{(n,i)'}\mathbf{ac}$, defined as

$$M^{(n,i)'}\mathbf{ac}=\min\left(M^{(n,i)}\mathbf{ac},M^{(n,i)}\mathbf{ab}+M^{(n,i)}\mathbf{bc}\right).$$

(Note that in every triangle $\mathbf{xyz}$ the triangle inequality is tested only in the form $D\mathbf{xz}\leq D\mathbf{xy}+D\mathbf{yz}$, irrespective of whether any of the remaining five triangle inequalities is violated, $D\mathbf{xy}\leq D\mathbf{xz}+D\mathbf{zy}$, $D\mathbf{zy}\leq D\mathbf{zx}+D\mathbf{xy}$, etc.)

The function $M^{(n,i)}$ for every (n,i) is clearly a dissimilarity function, and it is referred to as the *corrected dissimilarity function*. If, at some (n,i), the function $M^{(n,i)}$ is a quasimetric dissimilarity, it is called the *terminal corrected dissimilarity function*.

It follows from Definition 3.11 that if $(m,j)\geq(n,i)$, then $M^{(m,j)}\mathbf{xy}\leq M^{(n,i)}\mathbf{xy}$ for all $\mathbf{x},\mathbf{y}\in\mathfrak{S}$. Therefore, if, for some n, $M^{(n+1,t)}\equiv M^{(n,t)}$, then $M^{(n+1,1)}\equiv M^{(n,t)}$, implying that $M^{(n,t)}$ is the terminal dissimilarity function. The converse being obvious, we have

Lemma 3.12. *$M^{(n,i)}$ is the terminal corrected dissimilarity function if and only if $M^{(n+1,t)}\equiv M^{(n,t)}$.*

The next lemma provides a link between the algorithm being considered and the use of chains in the definition of G. Recall that $\mathcal{C}$ denotes the set of all chains in $\mathfrak{S}$.

Lemma 3.13. *For any $n=0,1,\ldots$, any $i=1,2,\ldots,t$, and any $\mathbf{a},\mathbf{b}\in\mathfrak{S}$, there is a chain $\mathbf{X}\in\mathcal{C}$ such that*

$$M^{(n,i)}\mathbf{ab}=D\mathbf{aXb}.$$

The proof obtains by induction on the lexicographically ordered (n,i). The statement holds for $n=0$, with $\mathbf{X}$ an empty chain. Let it hold for all double indices up to and including $(n,i)\geq(0,t)$, and let $\mathbf{abc}$ be the triangle indexed $(n,i)'$. Then the statement is clearly true for $M^{(n,i)'}\mathbf{ac}$ whether it equals $M^{(n,i)}\mathbf{ac}$ or $M^{(n,i)}\mathbf{ab}+M^{(n,i)}\mathbf{bc}$, and it is true for all other $\mathbf{xy}$ because then $M^{(n,i)'}\mathbf{xy}=M^{(n,i)}\mathbf{xy}$.

Does a terminal dissimilarity function necessarily exist? Let us assume it does not. Then, by Lemma 3.12, $M^{(n+1,t)}$ and $M^{(n,t)}$ do not coincide for all $n=0,1,\ldots$. Since $\mathfrak{S}\times\mathfrak{S}$ is finite, there should exist distinct points $\mathbf{a},\mathbf{b}\in\mathfrak{S}$ and an infinite sequence of positive integers $n_1<n_2<\ldots$ for which

$$D\mathbf{ab} \neq M^{(n_1,t)}\mathbf{ab} \neq M^{(n_2,t)}\mathbf{ab} \neq \dots.$$

From Definition 3.11 it follows then that

$$D\mathbf{ab} > M^{(n_1,t)}\mathbf{ab} > M^{(n_2,t)}\mathbf{ab} > \dots.$$

By Lemma 3.13, for every (n_i, t) there should exist a chain $\mathbf{X}_i$ such that

$$M^{(n_i,t)}\mathbf{ab} = D\mathbf{aX}_i\mathbf{b},\ i = 1, 2, \dots.$$

But a sequence of inequalities

$$D\mathbf{ab} > D\mathbf{aX}_{n_1}\mathbf{b} > D\mathbf{aX}_{n_2}\mathbf{b} > \dots$$

is impossible in a finite set, because the set of chains with lengths below a given value is finite. This contradiction proves the existence of a terminal dissimilarity function. Let us denote it by M. Observe that for any $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$ and any chain $\mathbf{X} \in \mathcal{C}$,

$$D\mathbf{aXb} \geq M\mathbf{aXb}.$$

But M satisfies the triangle inequality, whence

$$M\mathbf{aXb} \geq M\mathbf{ab},$$

whence

$$M\mathbf{ab} \leq D\mathbf{aXb}.$$

By Lemma 3.13, this implies

$$M\mathbf{ab} = \min_{\mathbf{X}\in\mathcal{C}} D\mathbf{aXb},$$

which equals $G\mathbf{ab}$ by definition. We have established therefore

Theorem 3.14. *A terminal corrected dissimilarity function exists, and it coincides with the quasimetric dissimilarity G induced by the initial dissimilarity function D.*

It is worthwhile emphasizing that nowhere in the proof have we used a specific order of the triangles in $\mathbf{T}^{(n)}$.

We see that dissimilarities on finite sets can be viewed as "imperfect" quasimetric dissimilarities, and the dissimilarity cumulation procedure can be recast as a series of recursive corrections of the dissimilarities for the violations of the triangle inequality.

Let us illustrate the procedure on our toy example, starting with the matrix of dissimilarities

$\Psi^{(1)} = D$	**a**	**b**	**c**	**d**
a	0	0.1	0.3	0.4
b	0.2	0	0.1	0.3
c	0.4	0.2	0	0.1
d	0.7	0.5	0.2	0

and using, for each $\mathbf{T}^{(n)}$ the same sequence of $t = 24$ triangles

$$\begin{array}{cccccc} i=1 & 2 & 3 & \dots & 23 & 24 \\ \mathbf{acb} & \mathbf{adb} & \mathbf{abc} & \dots & \mathbf{dac} & \mathbf{dbc} \end{array}. \tag{3.26}$$

It is obtained by cycling through the first element (4 values), subcycling through the last element (3 values), and sub-subcycling through the middle element (2 values), in alphabetic order.

Testing the triangles in $\mathbf{T}^{(1)}$ one by one, $M^{(1,1)}$coincides with D because the triangle indexed $(1,1)$ is **acb**, and the triangle inequality in it is not violated. Similarly, $M^{(1,2)} \equiv M^{(1,1)}$ because the triangle inequality is not violated in the triangle labeled $(1,2)$. The first violation of the triangle inequality occurs in the triangle indexed $(1,3)$, **abc**:

$$0.3 = D\mathbf{ac} > D\mathbf{ab} + D\mathbf{bc} = 0.1 + 0.1.$$

We "correct" the value of $D\mathbf{ac}$ therefore by replacing 0.3 with 0.2 (shown in parentheses in matrix $M^{(1,3)}$ below):

D	**a**	**b**	**c**	**d**
a	0	0.1	0.3	0.4
b	0.2	0	0.1	0.3
c	0.4	0.2	0	0.1
d	0.7	0.5	0.2	0

$\Rightarrow$

$M^{(1,3)}$	**a**	**b**	**c**	**d**
a	0	0.1	(0.2)	0.4
b	0.2	0	0.1	0.3
c	0.4	0.2	0.0	0.1
d	0.7	0.5	0.2	0.0

No violations occur until we reach the triangle indexed $(1,20)$, so $M^{(1,19)} \equiv M^{(1,18)} \equiv \dots \equiv M^{(1,3)}$. In $M^{(1,19)}$, however, we have, for the triangle **dca**:

$$0.7 = M^{(1,19)}\mathbf{da} > M^{(1,19)}\mathbf{dc} + M^{(1,19)}\mathbf{ca} = 0.2 + 0.4.$$

We correct $M^{(1,19)}\mathbf{da}$ from 0.7 to 0.6, as shown in the parentheses in matrix $M^{(1,20)}$:

$M^{(1,19)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.2	0.4
b	0.2	0	0.1	0.3
c	0.4	0.2	0	0.1
d	0.7	0.5	0.2	0

$\Rightarrow$

$M^{(1,20)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.2	0.4
b	0.2	0	0.1	0.3
c	0.4	0.2	0	0.1
d	(0.6)	0.5	0.2	0

We deal analogously with the third violation of the triangle inequality, in the triangle **dcb**, indexed $(1,22)$:

$$0.5 = M^{(1,21)}\mathbf{db} > M^{(1,21)}\mathbf{dc} + M^{(1,21)}\mathbf{cb} = 0.2 + 0.2.$$

So $M^{(1,21)} \equiv M^{(1,20)} \equiv M^{(1,19)}$, and

$M^{(1,21)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.2	0.4
b	0.2	0	0.1	0.3
c	0.4	20.	0	0.1
d	0.6	0.5	0.2	0

$\Rightarrow$

$M^{(1,22)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.2	0.4
b	0.2	0	0.1	0.3
c	0.4	20.	0	0.1
d	0.6	(0.4)	0.2	0

With the remaining two triangles before the sequence $\mathbf{T}^{(1)}$ has been exhausted no violations occur, so $M^{(1,24)} \equiv M^{(1,23)} \equiv M^{(1,22)}$ is the matrix with which the second sequence, $\mathbf{T}^{(2)}$, begins. The first and only violation here occurs at the triangle indexed $(2,5)$, **abd**:

$$0.4 = M^{(2,4)}\mathbf{ad} > M^{(2,4)}\mathbf{ab} + M^{(2,4)}\mathbf{bd} = 0.1 + 0.2.$$

So $M^{(2,4)} \equiv \cdots \equiv M^{(2,1)} \equiv M^{(1,24)}$, and

$M^{(1,24)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.2	0.4
b	0.2	0	0.1	0.3
c	0.4	0.2	0	0.1
d	0.6	0.4)	0.2	0

$\Rightarrow$

$M^{(2,5)}$	**a**	**b**	**c**	**d**
a	0	0.1	0.2	(0.3)
b	0.2	0	0.1	0.3
c	0.4	0.2	0	0.1
d	0.6	0.4	0.2	0

One can verify that $M^{(2,5)}$ is a quasimetric dissimilarity on $\mathfrak{S} = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$, so that $M^{(2,6)}$ and all higher-indexed matrices remain equal to $M^{(2,5)}$. The latter therefore is the terminal corrected dissimilarity, and its comparison with (3.24) shows that it coincides with $G = G^{(1)}$, the quasimetric induced by the initial dissimilarity function $D = \Psi^{(1)}$.

3.7 Dissimilarity Cumulation in Path-Connected Spaces

3.7.1 Chains-on-Nets and Paths

We now turn to dissimilarity cumulation in stimulus spaces $(\mathfrak{S}, D)$ in which points can be connected by *paths*. A path is a continuous function $\mathbf{f} : [a,b] \to \mathfrak{S}$. Because $[a,b]$ is a closed interval of reals, this function is also uniformly continuous. The latter means that $\mathbf{f}(x) \leftrightarrow \mathbf{f}(y)$ if $x - y \to 0$ $(x, y \in [a,b])$. We will present this path more compactly as $\mathbf{f}|\,[a,b]$, and say that it connects $\mathbf{f}(a) = \mathbf{a}$ to $\mathbf{f}(b) = \mathbf{b}$, where $\mathbf{a}$ and $\mathbf{b}$ are allowed to coincide.

To introduce the notion of the length of the path $\mathbf{f}|\,[a,b]$, we need the following auxiliary notions. A *net* on $[a,b]$ is defined as a sequence of numbers

$$\mu = (a = x_0 \le x_1 \le \cdots \le x_k \le x_{k+1} = b),$$

not necessarily pairwise distinct. The quantity

$$\delta\mu = \max_{i=0,1\ldots,k} (x_{i+1} - x_i)$$

is called the net's *mesh*. A net $\mu = (a, x_1, \ldots, x_k, b)$ can be elementwise paired with a chain $\mathbf{X} = \mathbf{x}_0\mathbf{x}_1 \ldots \mathbf{x}_k\mathbf{x}_{k+1}$ to form a *chain-on-net*

$$\mathbf{X}^{\mu} = ((a, \mathbf{x_0}), (x_1, \mathbf{x}_1), \ldots, (x_k, \mathbf{x}_k), (b, \mathbf{x}_{k+1})).$$

Note that the elements of the chain $\mathbf{X}$ need not be pairwise distinct. The *separation* of the chain-on-net $\mathbf{X}^{\mu}$ from the path $\mathbf{f}|\,[a,b]$ is defined as

$$\sigma\left(\mathbf{f}, \mathbf{X}^{\mu}\right) = \max_{x_i \in \mu} D\mathbf{f}(x_i)\,\mathbf{x}_i.$$

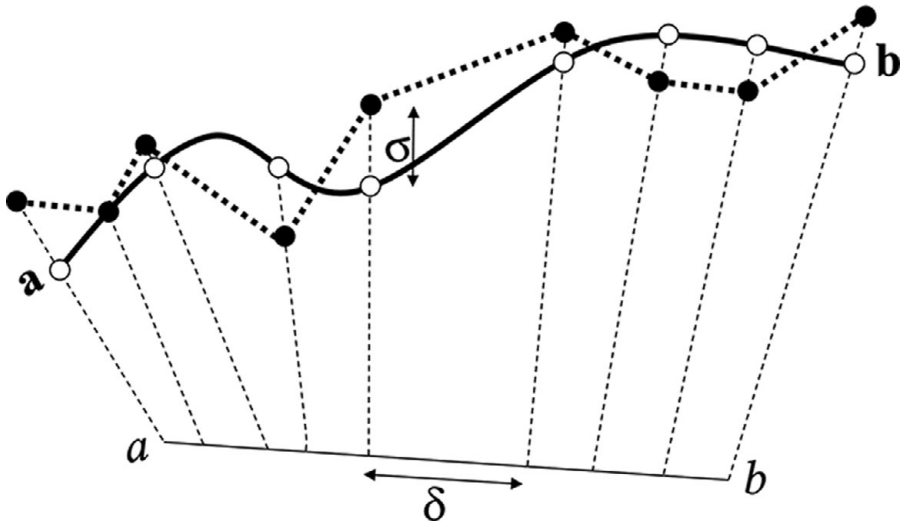


Figure 3.11 *Chains-on-nets* $\mathbf{X}^{\mu}$ *are converging to a path* $\mathbf{f}|\,[a,b]$ *as* $\delta = \delta\mu \to 0$ *and* $\sigma = \sigma\left(\mathbf{f}, \mathbf{X}^{\mu}\right) \to 0$. *The length* $D\mathbf{X}$ *of the chains then has the limit inferior that is taken for the length of the path* $\mathbf{f}$.

Definition 3.15. The *D-length* of path $\mathbf{f}|\,[a,b]$ is defined as

$$D\mathbf{f} = \liminf_{\delta\mu\to 0, \sigma(\mathbf{f},\mathbf{X}^{\mu})\to 0} D\mathbf{X}.$$

The limit inferior stands here for

$$\sup_{\varepsilon_1>0,\varepsilon_2>0} \inf\left\{D\mathbf{X} : \delta\mu < \varepsilon_1, \sigma\left(\mathbf{f}, \mathbf{X}^{\mu}\right) < \varepsilon_2\right\}.$$

Let us agree to say that $\mathbf{X}^{\mu}$ converges *to* $\mathbf{f}$ (and write $\mathbf{X}^{\mu} \to \mathbf{f}$) if $\delta\mu \to 0$ and $\sigma\left(\mathbf{f}, \mathbf{X}^{\mu}\right) \to 0$. We can then rewrite the definition above as

$$D\mathbf{f} = \liminf_{\mathbf{X}^{\mu}\to\mathbf{f}} D\mathbf{X}. \tag{3.27}$$

Using the properties of lim inf, for any path $\mathbf{f}$, there exists a sequence $\left\{\mathbf{X}_n^{\mu_n}\right\}$ of chains-on-nets such that $\delta\mu_n \to 0$ and $\sigma\left(\mathbf{f}, \mathbf{X}_n^{\mu_n}\right) \to 0$, and $D\mathbf{X}_n \to D\mathbf{f}$.

Let us list some of the most basic properties of the D-length of a path.

Theorem 3.16. *The length* $D\mathbf{f}$ *of any path* $\mathbf{f}|\,[a,b]$ *has the following properties:*

$\mathcal{L}1$ *(non-negativity)* $D\mathbf{f} \geq 0$;

$\mathcal{L}2$ *(zero property)* $D\mathbf{f} = 0$ *if and only if* $\mathbf{f}\left([a,b]\right)$ *is a single point;*

$\mathcal{L}3$ *(additivity) for any* $c \in [a,b]$, $D\mathbf{f}|\,[a,b] = D\mathbf{f}|\,[a,c] + D\mathbf{f}|\,[c,b]$.

Proofs of these statements are simple. Thus, to show the additivity of $D\mathbf{f}$, add the point c twice to all nets:

$$\widetilde{\mu} = \left\{\overbrace{a = x_0 \leq \ldots \leq x_i \leq c}^{\alpha} = \underbrace{c \leq x_{i+1} \leq \ldots \leq x_{k+1} = b}_{\beta}\right\},$$

and two corresponding points $\mathbf{c}^1, \mathbf{c}^2$ to all chains:

$$\widetilde{\mathbf{X}} = \overbrace{\mathbf{x}_0 \ldots \mathbf{x}_i \mathbf{c}^1}^{\mathbf{Y}} \underbrace{\mathbf{c}^2 \mathbf{x}_{i+1} \ldots \mathbf{x}_{k+1}}_{\mathbf{Z}}.$$

Clearly

$$\liminf_{\widetilde{\mathbf{X}}^{\widetilde{\mu}} \to \mathbf{f}|[a,b]} \widetilde{\mathbf{X}} = \liminf_{\mathbf{Y}^{\alpha} \to \mathbf{f}|[a,c]} \mathbf{Y} + \liminf_{\mathbf{Z}^{\beta} \to \mathbf{f}|[c,b]} \mathbf{Z} = D\mathbf{f}|\,[a,c] + D\mathbf{f}|\,[c,b]\,.$$

For any sequence $\left\{\mathbf{X}_n^{\mu_n}\right\}$ of chains-on-nets such that $\mathbf{X}_n^{\mu_n} \to \mathbf{f}|\,[a,c]$, and $D\mathbf{X}_n \to D\mathbf{f}$, we have $\widetilde{\boldsymbol{X}}_n^{\widetilde{\mu}_n} \to \mathbf{f}|\,[a,c]$ for the corresponding sequence $\left\{\widetilde{\boldsymbol{X}}_n^{\widetilde{\mu}_n}\right\}$, assuming $\mathbf{c}_n^1 \to \mathbf{f}\,(c)$ and $\mathbf{c}_n^2 \to \mathbf{f}\,(c)$. We also have

$$D\widetilde{\boldsymbol{X}}_n = D\mathbf{X}_n + \left(D\mathbf{x}_{i_n}\mathbf{c}_n^1 + D\mathbf{c}_n^1\mathbf{c}_n^2 + D\mathbf{c}_n^2\mathbf{x}_{i_n+1} - D\mathbf{x}_{i_n}\mathbf{x}_{i_n+1}\right),$$

where each summand in the parentheses tends to zero by the uniform continuity of $\mathbf{f}$ and D.

Note that $D\mathbf{f}$ is well-defined for any path $\mathbf{f}$, but only on the extended set of non-negative reals: the value of $D\mathbf{f}$ may very well be equal to ∞. This does not invalidate or complicate any of the results presented in this chapter, but, for brevity's sake, we will tacitly assume that $D\mathbf{f}$ is finite.

The reader may wonder why, in the definition of $D\mathbf{f}$, it is not sufficient to deal with the inscribed chains-on-nets, with all elements of the chains belonging to the path $\mathbf{f}$. We will see later that this is indeed sufficient if D is a quasimetric dissimilarity. However, in general, the inscribed chains-on-nets do not reach the infimum of the D-lengths of the "meandering" chains-on-nets. Figure 3.12 provides an illustration. In this example, the stimuli are points in $\mathbb{R}^2$, and, for $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$,

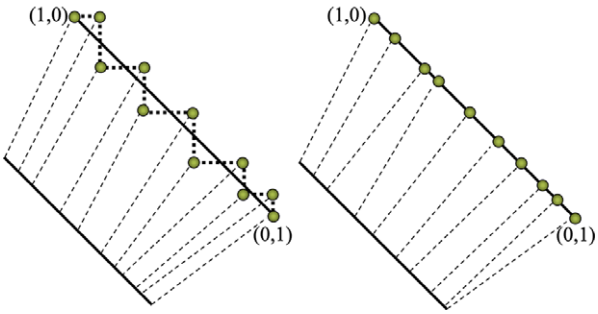


Figure 3.12 *A demonstration of why for D-length computations we need the "meandering" chains like in Figure 3.11 rather than just inscribed chains. Here, D***ab** *for* $\mathbf{a} = (a_1, a_2)$ *and* $\mathbf{b} = (b_1, b_2)$ *is defined as* $|a_1 - b_1| + |a_2 - b_2| + \min(|a_1 - b_1|, |a_2 - b_2|)$. *All staircase chains* $\mathbf{X}$*, irrespective of the spacing of their elements, have the cumulated dissimilarity* $D\mathbf{X} = 2$*, and* 2 *is the true D-length of the path between* $(1,0)$ *and* $(0,1)$. *All inscribed chains, irrespective of the spacing of their elements, have the cumulated dissimilarity* 3. *Explanations are given in the text.*

$$D\mathbf{ab} = |a_1 - b_1| + |a_2 - b_2| + \min\left(|a_1 - b_1|, |a_2 - b_2|\right).$$

It is easy to check that D is a dissimilarity function. Thus, $\mathcal{D}3$ follows from the fact

$$D\mathbf{a}_n\mathbf{b}_n \to 0 \Longleftrightarrow |\mathbf{a} - \mathbf{b}| \to 0,$$

where $|\mathbf{a} - \mathbf{b}|$ is the usual Euclidean norm. Also, for any chain $\mathbf{aXb}$,

$$D\mathbf{aXb} \geq |a_1 - b_1| + |a_2 - b_2|,$$

whence $D\mathbf{a}_n\mathbf{X}_n\mathbf{b}_n \to 0$ implies $D\mathbf{a}_n\mathbf{b}_n \to 0$. That is, D satisfies $\mathcal{D}4$. By the same inequality, the length of the line segment $\mathbf{f}$ shown in Figure 3.12, connecting $\mathbf{a} = (1,0)$ to $\mathbf{b} = (0,1)$, cannot be less than 2. (The domain interval for $\mathbf{f}$ can be chosen arbitrarily, e.g., $[0,1]$.) Consider now chains-on-nets $\mathbf{X}^\mu$ with the staircase chains, as in the left panel. By decreasing the mesh of μ and the spacing of the elements of $\mathbf{X}$, it can be made to converge to $\mathbf{f}$, and since $D\mathbf{X}$ for all these chains equals 2, $D\mathbf{f} = 2$. At the same time, the inscribed chains, as in the right panel of the figure, are easily checked to have the length 3.

3.7.2 Path Length through Quasimetric Dissimilarity

Different dissimilarity functions D lead to different quantifications of path length. We know that the quasimetric dissimilarity G defined by (3.18) is a dissimilarity function. However, in this case, since G is defined through D by (3.18), one should expect, for consistency, that the path length will remain unchanged on replacing D with G. This will indeed be established in Section 3.7.3. We need several preliminary results first, however.

Using G in place of D to define the G-length of paths, we have

$$G\mathbf{f} = \liminf_{\mathbf{X}^\mu \overset{G}{\to} \mathbf{f}} G\mathbf{X}.$$

The condition $\mathbf{X}^\mu \overset{G}{\to} \mathbf{f}$ here means $\delta\mu \to 0$ and

$$\sigma_G\left(\mathbf{f}, \mathbf{X}^\mu\right) = \max_{x_i \in \mu} G\mathbf{f}\left(x_i\right)\mathbf{x}_i \to 0.$$

But, by Theorem 3.6, the latter condition is equivalent to

$$\sigma\left(\mathbf{f}, \mathbf{X}^\mu\right) = \max_{x_i \in \mu} D\mathbf{f}\left(x_i\right)\mathbf{x}_i \to 0.$$

Therefore $\mathbf{X}^\mu \overset{G}{\to} \mathbf{f}$ and $\mathbf{X}^\mu \to \mathbf{f}$ are equivalent, and we can formulate

Definition 3.17. The *G-length* of path $\mathbf{f}|\,[a,b]$ is

$$G\mathbf{f} = \liminf_{\mathbf{X}^\mu \to \mathbf{f}} G\mathbf{X}.$$

Consider now chains-on-nets $\mathbf{Z}^\nu$ *inscribed* in $\mathbf{f}|\,[a,b]$, that is, those with

$$\nu = \{a = z_0, z_1, \ldots, z_k, z_{k+1} = b\}$$

and

$$\mathbf{Z} = \mathbf{f}(z_0)\ldots\mathbf{f}(z_{k+1}) = \mathbf{z}_0\ldots\mathbf{z}_{k+1}.$$

Since $\sigma(\mathbf{f}, \mathbf{Z}^\nu) = 0$, the condition $\mathbf{Z}^\nu \to \mathbf{f}$ here reduces to $\delta\nu \to 0$. Clearly

$$\liminf_{\delta\nu\to 0} G\mathbf{Z} \geq \liminf_{\mathbf{X}^\mu \to \mathbf{f}} G\mathbf{X} = G\mathbf{f}, \tag{3.28}$$

because inscribed chains-on-nets converging to $\mathbf{f}$ form a subset of all chains-on-nets converging to $\mathbf{f}$. We will see now that in fact the two quantities in (3.28) are equal. By the additivity property,

$$G\mathbf{f}|[a,b] = \sum_{i=0}^{k} G\mathbf{f}|\left[z_i, z_{i+1}\right].$$

Let $\mathbf{X}_i^{\mu_i}$ be an arbitrary chain-on-net with $\mu_i \subset \left[z_i, z_{i+1}\right]$. By the same reasoning as in the proof of the additivity property, if μ_i is changed into

$$\widetilde{\mu}_i = \left\{z_i, \overbrace{\ldots}^{\mu_i}, z_{i+1}\right\}$$

and $\mathbf{X}_i$ into

$$\widetilde{\mathbf{X}}_i = \mathbf{z}_i \mathbf{X}_i \mathbf{z}_{i+1},$$

the conditions $\mathbf{X}_i^{\mu_i} \to \mathbf{f}|\left[z_i, z_{i+1}\right]$ and $\widetilde{\boldsymbol{X}}_i^{\widetilde{\mu}_i} \to \mathbf{f}|\left[z_i, z_{i+1}\right]$ are equivalent. Denoting by $\mathbf{X}^\mu$ the concatenation of $\mathbf{X}_i^{\mu_i}$ for $i = 0, \ldots, k$, and defining $\widetilde{\boldsymbol{X}}^{\widetilde{\mu}}$ analogously, we have

$$G\mathbf{f} = \liminf_{\mathbf{X}^\mu \to \mathbf{f}} G\mathbf{X} = \liminf_{\widetilde{\boldsymbol{X}}^{\widetilde{\mu}} \to \mathbf{f}} G\widetilde{\boldsymbol{X}}.$$

At the same time, by the triangle inequality,

$$G\mathbf{z}_i\mathbf{z}_{i+1} \leq G\mathbf{z}_i\mathbf{X}_i\mathbf{z}_{i+1},$$

whence

$$G\mathbf{Z} \leq G\widetilde{\boldsymbol{X}}$$

and

$$\liminf_{\delta\nu\to 0} G\mathbf{Z} \leq \liminf_{\widetilde{\boldsymbol{X}}^{\widetilde{\mu}} \to \mathbf{f}} G\widetilde{\boldsymbol{X}} = G\mathbf{f}. \tag{3.29}$$

Together with (3.28), this establishes

Theorem 3.18. *For any path* $\mathbf{f}$*,*

$$G\mathbf{f} = \liminf_{\delta\nu\to 0} G\mathbf{Z},$$

where $\mathbf{Z}^\nu$ *are chains-on-nets inscribed in* $\mathbf{f}$*.*

In other words, to approximate $G\mathbf{f}$ by G-lengths of chains-on-nets, one does not need all possible chains converging to $\mathbf{f}$; the inscribed ones only are sufficient. Recall that the analogous statement is not correct for $D\mathbf{f}$. The equality in Theorem 3.18 critically owes to the fact that G satisfies the triangle inequality.

We can further clarify Theorem 3.18 as follows.

Theorem 3.19. *For any path* $\mathbf{f}$*,*

$$G\mathbf{f} = \sup G\mathbf{Z} = \lim_{\delta\nu\to 0} G\mathbf{Z}, \tag{3.30}$$

where $\mathbf{Z}^{\nu}$ *are chains-on-nets inscribed in* $\mathbf{f}$*.*

In other words, $G\mathbf{f}$ is the lowest upper bound for the lengths of all inscribed chains-on-nets; and any sequence of the inscribed chains-on-nets converges to $G\mathbf{f}$ as their mesh decreases.

To prove the first equality, $G\mathbf{f} = \sup G\mathbf{Z}$, consider a chain-on-net $\mathbf{Z}^{\nu}$ with $\sup G\mathbf{Z} - G\mathbf{Z}$ arbitrarily small. For every pair of successive z_i, z_{i+1} in ν, one can find an inscribed chain-on-net $\mathbf{V}_i^{\mu_i}$ such that $\mu_i = \{z_i, \ldots, z_{i+1}\}$ and $\left|G\mathbf{V}_i - G\mathbf{f}\right|\left[z_i, z_{i+1}\right]\Big|$ is arbitrarily small. By the additivity of G-length, denoting by $\mathbf{V}^{\mu}$ the concatenation of all $\mathbf{V}_i^{\mu_i}$, we can make $|G\mathbf{V} - G\mathbf{f}|\,[a,b]|$ arbitrarily small. From the triangle inequality it follows that $G\mathbf{V} \geq G\mathbf{Z}$, whence $G\mathbf{f} \geq \sup G\mathbf{Z}$. But $G\mathbf{V} \leq \sup G\mathbf{Z}$, whence we also have $G\mathbf{f} \leq \sup G\mathbf{Z}$.

To prove that $G\mathbf{f} = \lim_{\delta\nu\to 0} G\mathbf{Z}$, deny it, and assume that there is a sequence of inscribed chains-on-nets $\mathbf{V}_n^{\mu_n}$ such that $\delta\mu_n \to 0$ but $G\mathbf{V}_n \not\to G\mathbf{f}$. Since $G\mathbf{f} = \sup G\mathbf{Z}$ across all possible inscribed chains-on-nets, $G\mathbf{V}_n \leq G\mathbf{f}$ for all n. Then one can find a $\Delta > 0$ and a subsequence of $\mathbf{V}_n^{\mu_n}$ (which, with no loss of generality, we can assume to be $\mathbf{V}_n^{\mu_n}$ itself) such that

$$G\mathbf{V}_n \to G\mathbf{f} - \Delta.$$

Let $\mathbf{Z}^{\nu}$ be an inscribed chain-on-net with

$$G\mathbf{Z} > G\mathbf{f} - \Delta/2.$$

For every z_i in ν and every n, let $v_{k_{i,n}}^n, v_{k_{i,n}+1}^n$ be two successive elements of μ_n such that $v_{k_{i,n}}^n \leq z_i \leq v_{k_{i,n}+1}^n$. For a sufficiently large n, $\delta\mu_n$ is sufficiently small to ensure that z_i is the only member of ν falling between $v_{k_{i,n}}^n$ and $v_{k_{i,n}+1}^n$ (without loss of generality, we can assume that ν contains no identical elements). Denote by $\nu \uplus \mu_n$ the nets formed by the elements of ν inserted into μ_n. Consider the inscribed chains-on-nets $\mathbf{U}^{\nu\uplus\mu_n}$. We have (denoting by l the cardinality of ν)

$$G\mathbf{U} = G\mathbf{V}_n + \sum_{i=0}^{l}\left\{G\mathbf{f}\left(v_{k_{i,n}}^n\right)\mathbf{f}\left(z_i\right) + G\mathbf{f}\left(z_i\right)\mathbf{f}\left(v_{k_{i,n}+1}^n\right) - G\mathbf{f}\left(v_{k_{i,n}}^n\right)\mathbf{f}\left(v_{k_{i,n}+1}^n\right)\right\}.$$

By the uniform continuity of $\mathbf{f}$, the expression under the summation operator tends to zero, whence

$$G\mathbf{U} - G\mathbf{V}_n \to 0,$$

and then

$$G\mathbf{U} \to G\mathbf{f} - \Delta.$$

But by the triangle inequality, for all n,

$$G\mathbf{U} \geq G\mathbf{Z} > G\mathbf{f} - \Delta/2.$$

This contradiction completes the proof.

3.7.3 The Equality of the D-length and G-length of Paths

As mentioned previously, one can expect that path length should not depend on whether one chooses dissimilarity D or the quasimetric dissimilarity G induced by D.

Theorem 3.20. *For any path* $\mathbf{f}$*,*

$$D\mathbf{f} = G\mathbf{f}.$$

Comparing Definitions 3.15 and 3.17, since $D\mathbf{X} \geq G\mathbf{X}$ for any chain, we have $D\mathbf{f} \geq G\mathbf{f}$. To see that $D\mathbf{f} \leq G\mathbf{f}$, we form a sequence of inscribed chains-on-nets $\mathbf{Z}_n^{\nu_n}$ such that $\delta\nu_n \to 0$, and $G\mathbf{Z}_n \to G\mathbf{f}$. By the definition of G, one can insert chains $\mathbf{X}_i^n$ between pairs of successive elements $\mathbf{z}_i^n, \mathbf{z}_{i+1}^n$ of $\mathbf{Z}_n$, so that

$$D\mathbf{U}_n - G\mathbf{Z}_n \leq \frac{1}{n},$$

where

$$\mathbf{U}_n = \mathbf{z}_0^n \mathbf{X}_0^n \mathbf{z}_1^n \dots \mathbf{z}_{k_n}^n \mathbf{X}_{k_n}^n \mathbf{z}_{k_n+1}^n.$$

In other words, $D\mathbf{U}_n \to G\mathbf{f}$. Let us now create a net μ_n for every $\mathbf{U}_n$ as follows: if $z_i^n \in \nu_n$ is associated with $\mathbf{z}_i^n \in \mathbf{Z}_n$, we associate z_i^n with every element of $\mathbf{X}_i^n$. The resulting chain-on-net is

$$\mathbf{U}_n^{\mu_n} = \left(\dots, \left(z_i^n, \mathbf{z}_i^n\right), \left(z_i^n, \mathbf{x}_1^{i,n}\right), \dots, \left(z_i^n, \mathbf{x}_{l_{i,n}}^{i,n}\right) \left(z_{i+1}^n, \mathbf{z}_{i+1}^n\right), \dots \right).$$

We will show now that $\mathbf{U}_n^{\mu_n} \to \mathbf{f}$. Since $\delta\mu_n = \delta\nu_n \to 0$, we have to show that $\sigma\left(\mathbf{f}, \mathbf{U}_n^{\mu_n}\right) \to 0$. Let $\left(z_{i_n}^n, \mathbf{m}_{i_n}^n\right)$ be an element of $\mathbf{U}_n^{\mu_n}$ such that

$$\sigma\left(\mathbf{f}, \mathbf{U}_n^{\mu_n}\right) = D\mathbf{f}\left(z_{i_n}^n\right) \mathbf{m}_{i_n}^n = D\mathbf{z}_{i_n}^n \mathbf{m}_{i_n}^n.$$

By the uniform continuity of $\mathbf{f}$ and G,

$$G\mathbf{z}_{i_n}^n \mathbf{z}_{i_n+1}^n = G\mathbf{f}\left(z_{i_n}^n\right) \mathbf{f}\left(z_{i_n+1}^n\right) \to 0$$

as $\delta\mu_n = \delta\nu_n \to 0$. By the construction of $\mathbf{U}_n$,

$$D\mathbf{z}_{i_n}^n = D\mathbf{z}_{i_n}^n \overbrace{\mathbf{x}_1^{i_n,n} \dots \mathbf{m}_{i_n}^n \dots \mathbf{x}_{l_{i_n,n}}^{i_n,n}}^{\mathbf{X}_{i_n}^n} \mathbf{z}_{i_n+1}^n \to 0,$$

implying

$$D\mathbf{z}_{i_n}^n \mathbf{x}_1^{i_n,n} \dots \mathbf{m}_{i_n}^n \to 0.$$

By the chain property of dissimilarity functions,

$$\sigma\left(\mathbf{f}, \mathbf{U}_n^{\mu_n}\right) = D\mathbf{z}_{i_n}^n \mathbf{m}_{i_n}^n \to 0.$$

We have therefore a sequence of chains-on-nets $\mathbf{U}_n^{\mu_n} \to \mathbf{f}$ with $G\mathbf{f}$ as the limit point of $D\mathbf{U}_n$, and then $G\mathbf{f} \geq D\mathbf{f}$ because $D\mathbf{f}$ is the infimum of all such limit points. This completes the proof.

We see that although $D\mathbf{xy}$ and $G\mathbf{xy}$ are generally distinct for points $\mathbf{x}, \mathbf{y}$, when it comes to paths $\mathbf{f}$, the quantities $D\mathbf{f}$ and $G\mathbf{f}$ can be used interchangeably. One consequence of this result is that the properties of the D-length of paths can now be established by replacing it with the G-length, the advantage of this being that we acquire the powerful triangle inequality to use, and also restrict, chains-on-nets to the inscribed ones, more familiar than the "meandering" chains in Figure 3.11. However, the general definition of $D\mathbf{f}$ remains convenient in many situations. We illustrate this on the important property of *lower semicontinuity* of the D-length.

Definition 3.21. A sequence of paths $\mathbf{f}_n|\,[a,b]$ *converges* to a path $\mathbf{f}|\,[a,b]$ (in symbols, $\mathbf{f}_n \to \mathbf{f}$) if

$$\sigma\,(\mathbf{f},\mathbf{f}_n) = \max_{x\in[a,b]} D\mathbf{f}\,(x)\,\mathbf{f}_n\,(x) \to 0.$$

Consider any sequence of chains-on-nets $\mathbf{X}_n^{\mu_n} \to \mathbf{f}_n$ such that $|D\mathbf{X}_n - D\mathbf{f}_n| \to 0$. By the uniform continuity of D,

$$\left[\sigma\left(\mathbf{f}_n, \mathbf{X}_n^{\mu_n}\right) \to 0\right] \text{ and } [\sigma\,(\mathbf{f},\mathbf{f}_n) \to 0] \Longrightarrow \sigma\left(\mathbf{f}, \mathbf{X}_n^{\mu_n}\right) \to 0.$$

Then $\mathbf{X}_n^{\mu_n} \to \mathbf{f}$, whence $\liminf_{n\to\infty} D\mathbf{X}_n \geq D\mathbf{f}$. But $\liminf_{n\to\infty} D\mathbf{X}_n = \liminf_{n\to\infty} D\mathbf{f}_n$. This proves

Theorem 3.22 (Lower semicontinuity) *For any sequence of paths* $\mathbf{f}_n|\,[a,b] \to \mathbf{f}|\,[a,b]$,

$$\liminf_{n\to\infty} D\mathbf{f}_n \geq D\mathbf{f}.$$

3.7.4 Intrinsic Metrics and Spaces with Intermediate Points

In a path-connected space, a metric is traditionally called *intrinsic* if the distance between two points is the greatest lower bound for the length of all paths connecting the two points. For instance, in $\mathbb{R}^n$ endowed with the Euclidean geometry, the Euclidean distance

$$D\mathbf{ab} = |\mathbf{a} - \mathbf{b}|$$

between points $\mathbf{a}$ and $\mathbf{b}$ is intrinsic, because it is also the length of the shortest path connecting these points, a straight line segment. By contrast

$$D\mathbf{ab} = \sqrt{|\mathbf{a} - \mathbf{b}|}$$

is also a metric, but it is not intrinsic: the path length $D\mathbf{f}$ induced by this metric is infinitely large for every path $\mathbf{f}$. As an example of a non-intrinsic metric with a finite path length function, consider

$$D\mathbf{ab} = \tan|a - b|$$

on the interval $\left[0, \frac{\pi}{2}\right[$, where a, b are the values of $\mathbf{a}, \mathbf{b}$, respectively. The length of the (only) path connecting $\mathbf{a}$ to $\mathbf{b}$ here is $|a - b| \neq \tan|a - b|$.

In this section we consider a generalization of the notion of intrinsic metric to quasimetric dissimilarities.

Definition 3.23. The quasimetric dissimilarity G defined in a space $(\mathfrak{S}, D)$ by (3.18) is called intrinsic if, for any $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$,

$$G\mathbf{ab} = \inf_{\mathbf{f} \in \mathcal{P}_{\mathbf{a}}^{\mathbf{b}}} D\mathbf{f},$$

where $\mathcal{P}_{\mathbf{a}}^{\mathbf{b}}$ is the class of all *paths* connecting $\mathbf{a}$ to $\mathbf{b}$.

Figure 3.13 provides an illustration.

We know that in Definition 3.23 $D\mathbf{f}$ can be replaced with $G\mathbf{f}$. We also know that $G\mathbf{f}$ for any $\mathbf{f} \in \mathcal{P}_{\mathbf{a}}^{\mathbf{b}}$ can be arbitrarily closely approximated by $G\mathbf{aXb}$ for some inscribed chain-on-net $\mathbf{X}^{\mu}$. By the triangle inequality, $G\mathbf{ab} \leq G\mathbf{aXb}$. Therefore, in any space $(\mathfrak{S}, D)$,

$$G\mathbf{ab} \leq \inf_{\mathbf{f} \in \mathcal{P}_{\mathbf{a}}^{\mathbf{b}}} D\mathbf{f}. \tag{3.31}$$

We now need to consider a special class of spaces in which this inequality can be reversed.

Definition 3.24. A stimulus space $(\mathfrak{S}, D)$ is said to be a space *with intermediate points* if, for any distinct $\mathbf{a}, \mathbf{b}$, one can find an $\mathbf{m}$ such that $\mathbf{m} \notin \{\mathbf{a}, \mathbf{b}\}$ and $D\mathbf{amb} \leq D\mathbf{ab}$.

Figure 3.14 provides an illustration. If D is a metric (or quasimetric dissimilarity), the inequality $D\mathbf{amb} \leq D\mathbf{ab}$ can only have the form

$$D\mathbf{amb} = D\mathbf{ab}.$$

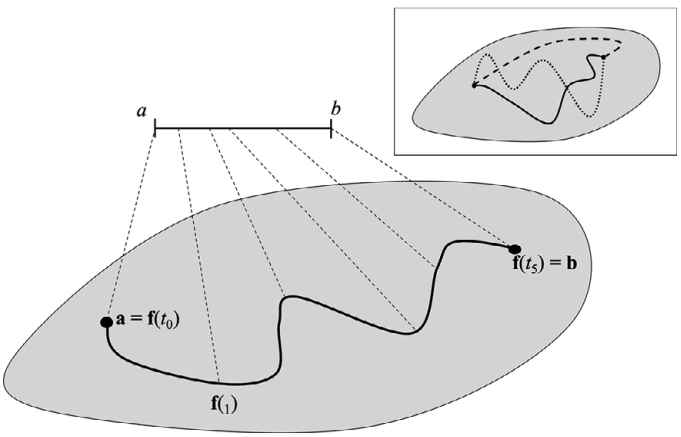


Figure 3.13 *The metric G induced by dissimilarity D is intrinsic if the G-distance from* **a** *to* **b** *equals the infimum of D-lengths (equivalently, G-lengths) of all paths connecting* **a** *to* **b**.

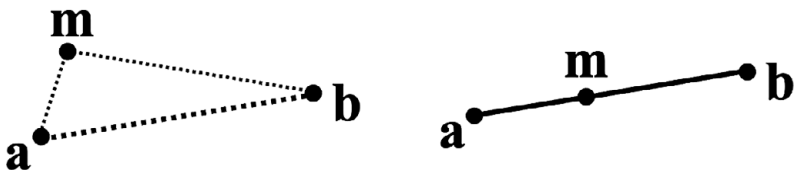


Figure 3.14 *If D***amb** $\leq$ *D***ab**, *the point* **m** *is said to be intermediate to* **a** *and* **b**. *As a special case, if D is Euclidean distance (right picture), any* **m** *on the straight-line segment connecting* **a** *and* **b** *is intermediate to* **a** *and* **b**.

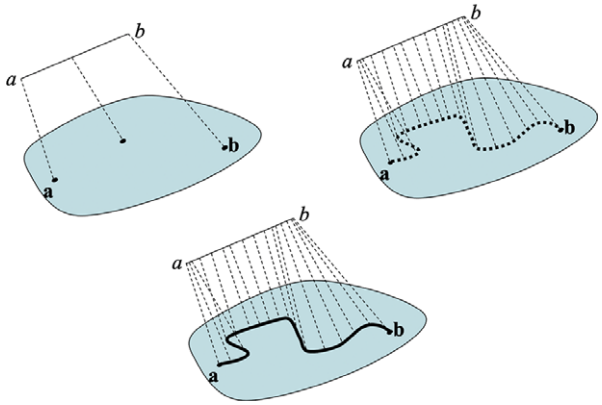


Figure 3.15 *An informal illustration of Theorems 3.26 and 3.27: by adding intermediate points for every pair of successive points one can create at the limit a path connecting* **a** *to* **b**, *with its D-length not exceeding D***ab**. *The infimum of the D-lengths of all such paths equals G***ab**.

In this form the notion is known as *Menger convexity*.

A sequence $\mathbf{x_1}, \mathbf{x}_2, \ldots$ in $(\mathfrak{S}, D)$ is called a *Cauchy sequence* if

$$\lim_{\substack{k\to\infty \\ l\to\infty}} D\mathbf{x}_k\mathbf{x}_l = 0,$$

that is, if for any $\varepsilon > 0$ one can find an n such that $D\mathbf{x}_k\mathbf{x}_l < \varepsilon$ whenever $k, l > n$.

Definition 3.25. A space $(\mathfrak{S}, D)$ is called *D-complete* (or simply, complete) if every Cauchy sequence in it converges to a point.

That is, in a complete space, for any Cauchy sequence $\mathbf{x_1}, \mathbf{x}_2, \ldots$, there is a point $\mathbf{x} \in \mathfrak{S}$ such that $\mathbf{x}_n \to \mathbf{x}$. For example, if stimuli are represented by points in a closed region of $\mathbb{R}^n$, and the convergence $\mathbf{x}_n \to \mathbf{x}$ coincides with the usual convergence of n-element vectors, then the space is complete.

The main mathematical fact we are interested in is as follows.

Theorem 3.26. *In a complete space* $(\mathfrak{S}, D)$ *with intermediate points, any point* **a** *can be connected to any point* **b** *by a path* **f** *with*

$$D\mathbf{f} \leq D\mathbf{ab}.$$

A proof of this statement known to us is rather involved (see Section 3.10 for a reference), and we will omit it here. Figure 3.15 provides an intuitive illustration.

A consequence of this theorem that is of special importance for us is as follows. In any sequence of chains $\mathbf{X}_n$ connecting $\mathbf{a}$ to $\mathbf{b}$, with $D\mathbf{X}_n \to G\mathbf{ab}$, each link $\mathbf{x}_{i_n}\mathbf{x}_{i_n+1}$ in each chain $\mathbf{X}_n$ can be replaced with a path $\mathbf{f}_{i_n}$ connecting $\mathbf{x}_{i_n}$ to $\mathbf{x}_{i_n+1}$, such that $D\mathbf{f}_{i_n} \leq D\mathbf{x}_{i_n}\mathbf{x}_{i_n+1}$. This would create a path $\mathbf{f}_n$ connecting $\mathbf{a}$ to $\mathbf{b}$, with $D\mathbf{f}_n \leq D\mathbf{X}_n$. Hence

$$\inf_{\mathbf{f}\in\mathcal{P}_{\mathbf{a}}^{\mathbf{b}}} D\mathbf{f} \leq \liminf_{n\to\infty} D\mathbf{f}_n \leq \lim_{n\to\infty} D\mathbf{X}_n = G\mathbf{ab}. \tag{3.32}$$

Combining this with (3.31), we establish

Theorem 3.27. *In a complete space* $(\mathfrak{S}, D)$ *with intermediate points, the quasi-metric dissimilarity G is intrinsic:*

$$G\mathbf{ab} = \inf_{\mathbf{f}\in\mathcal{P}_{\mathbf{a}}^{\mathbf{b}}} D\mathbf{f}.$$

3.8 Dissimilarity Cumulation in Euclidean Spaces

3.8.1 Introduction

We are now prepared to see how the general theory of path length can be specialized to a variant of (Finsler) differential geometry. We assume that in the canonical space of stimuli $(\mathfrak{S}, D)$, the set $\mathfrak{S}$ is an *open connected* region of the Euclidean n-space $\mathbb{R}^n$. The Euclidean n-space is endowed with the global coordinate system

$$\mathbf{x} = \left(x^1, \ldots, x^n\right),$$

and the conventional metric

$$E\mathbf{ab} = |\mathbf{a} - \mathbf{b}|. \tag{3.33}$$

Recall that the connectedness of $\mathfrak{S}$ means that it cannot be presented as a union of two open nonempty sets. In the Euclidean space this notion is equivalent to *path-connectedness*: any two points can be connected by a path.

Among all paths we focus on continuously differentiable ones. We develop a way of measuring the value $\widehat{F}\left(\mathbf{f}(x), \dot{\mathbf{f}}(x)\right)$ of the tangent vector $\dot{\mathbf{f}}(x)$ to the path $\mathbf{f}|[a,b]$ at point x, by showing (under certain assumptions) that

$$\widehat{F}\left(\mathbf{f}(x), \dot{\mathbf{f}}(x)\right) = \lim_{s\to 0+} \frac{G\mathbf{f}(x)\,\mathbf{f}(x+s)}{s}.$$

The D-length of the path is then computed as

$$\int_a^b \widehat{F}\left(\mathbf{f}(x), \dot{\mathbf{f}}(x)\right) dx.$$

The idea is illustrated in Figure 3.16.

We begin now a systematic development.

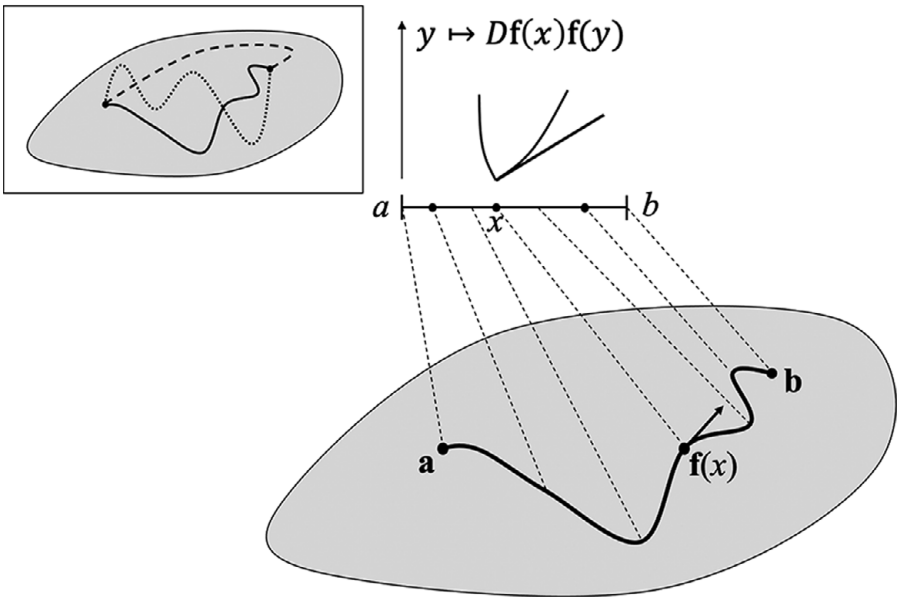


Figure 3.16 *As the point on the path moves away from a position* $\mathbf{f}(x)$*, the dissimilarity* $D\mathbf{f}(x)\,\mathbf{f}(y)$ *increases from zero, and the rate of this increase,* $dD\mathbf{f}(x)\,\mathbf{f}(x+s)\,/ds|_{s=0+}$ *is shown by the slope of the tangent line in the graph of* $y \mapsto D\mathbf{f}(x)\,\mathbf{f}(y)$*. This derivative then is integrated with respect to x from a to b to obtain the length of the path* $\mathbf{f}$*. If this derivative only depends on* $\mathbf{f}(x)$ *and* $d\mathbf{f}(x)\,/dx$ *(assuming the path is continuously differentiable), then it can be viewed as a way of measuring the tangent vector to the path as a point moves along it,* $F(\mathbf{f}(x), d\mathbf{f}(x)\,/dx)$*. The infimum of the lengths of all such smooth paths connecting* $\mathbf{a}$ *to* $\mathbf{b}$ *is then taken for the value of* $G\mathbf{ab}$*.*

Definition 3.28. The *tangent space* $\mathbb{T}_{\mathbf{p}}$ at a point $\mathbf{p}$ of $\mathfrak{S}$ is the set $\{\mathbf{p}\} \times \mathbb{U}^n$, where $\mathbb{U}^n$ is the *vector space*

$$\left\{\mathbf{u} = \mathbf{x} - \mathbf{p} : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{p}\right\}$$

endowed with the Euclidean vector norm $|\mathbf{u}|$ and the standard topology. The n-vectors $\mathbf{u} \in \mathbb{U}^n$ are referred to as *directions*, and the elements $(\mathbf{p}, \mathbf{u})$ of $\mathbb{T}_{\mathbf{p}}$ as *line elements*. The set of all line elements

$$\mathbb{T} = \mathfrak{S} \times \mathbb{U}^n = \bigcup_{\mathbf{p} \in \mathfrak{S}} \mathbb{T}_{\mathbf{p}}$$

is called the *tangent bundle* of the space $\mathfrak{S}$.

This definition deviates from the traditional one, which does not include the point $\mathbf{p}$ explicitly, but it is more convenient for our purposes. In the more general case of a *differentiable manifold*, the vector space $\mathbb{U}^n$ should be redefined. Note that the vectors in $\mathbb{U}^n$ do not represent stimuli, but we still use boldface letters to denote them. In the context of Euclidean spaces the boldface notation for both stimuli and directions can simply be taken as indicating vectors.

For any $\mathbf{u} \in \mathbb{U}^n$ the notation $\overline{\mathbf{u}}$ will be used for the unit vector codirectional with $\mathbf{u}$:

$$\overline{\mathbf{u}} = \frac{\mathbf{u}}{|\mathbf{u}|}, \qquad |\overline{\mathbf{u}}| = 1. \tag{3.34}$$

3.8.2 Submetric Function

We make the following two assumptions about the space $(\mathfrak{S}, D)$ and its relation to $(\mathfrak{S}, E)$.

($\mathcal{E}1$) The topologies of $(\mathfrak{S}, D)$ and $(\mathfrak{S}, E)$ coincide.

The coincidence of the D-topology and the Euclidean topology means that the notion of convergence

$$\mathbf{a}_n \to \mathbf{a} \tag{3.35}$$

means simultaneously $D\mathbf{a}_n\mathbf{a} \to 0$ and $|\mathbf{a}_n - \mathbf{a}| \to 0$. As a result, all topological concepts (openness, continuity, compactness, etc.) can be used without the prefixes D, G, or E. In particular, dissimilarity $D\mathbf{xy}$ and metric $G\mathbf{xy}$ are continuous in $(\mathbf{x}, \mathbf{y})$ with respect to the usual Euclidean topology.

Note, however, that the notions of uniform convergence in $(\mathfrak{S}, D)$ and $(\mathfrak{S}, E)$ are not assumed to coincide. Thus, it is possible that $D\mathbf{a}_n\mathbf{b}_n \to 0$ but $|\mathbf{a}_n - \mathbf{b}_n| \not\to 0$, or vice versa. In particular, dissimilarity $D\mathbf{xy}$ and metric $G\mathbf{xy}$ are not generally uniformly continuous in the Euclidean sense.

($\mathcal{E}2$) For any $\mathbf{x}, \mathbf{a}_n, \mathbf{b}_n \in \mathfrak{S}$ ($\mathbf{a}_n \neq \mathbf{b}_n$) and any unit vector $\overline{\mathbf{u}}$, if $\mathbf{a}_n \to \mathbf{x}$, $\mathbf{b}_n \to \mathbf{x}$, and $\overline{\mathbf{b}_n - \mathbf{a}_n} \to \overline{\mathbf{u}}$ (see Figure 3.17), then

$$\frac{D\mathbf{a}_n\mathbf{b}_n}{|\mathbf{b}_n - \mathbf{a}_n|}$$

tends to a positive limit, denoted $F(\mathbf{x}, \overline{\mathbf{u}})$.

Putting $\mathbf{a}_n = \mathbf{x}$ and $\overline{\mathbf{b}_n - \mathbf{a}_n} = \overline{\mathbf{u}}$ in Assumption $\mathcal{E}2$, and denoting $\mathbf{b}_n = \mathbf{x} + \overline{\mathbf{u}}s$, the function $F(\mathbf{x}, \overline{\mathbf{u}})$ can be presented as

$$F(\mathbf{x}, \overline{\mathbf{u}}) = \lim_{s \to 0+} \frac{D\mathbf{x}\,[\mathbf{x} + \overline{\mathbf{u}}s]}{s}. \tag{3.36}$$

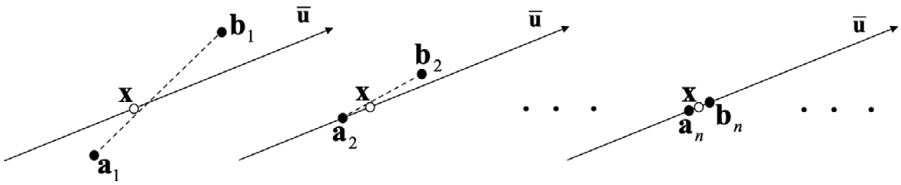


Figure 3.17 *An illustration for Assumption $\mathcal{E}2$. Shown are a point $\mathbf{x}$ (open circle), a direction $\overline{\mathbf{u}}$ attached to it, and (in successive panels from left to right) pairs of points $(\mathbf{a}_1, \mathbf{b}_1)$, $(\mathbf{a}_2, \mathbf{b}_2)$, ..., $(\mathbf{a}_n, \mathbf{b}_n)$, ... gradually converging to $\mathbf{x}$ so that the dashed line connecting them (and directed from $\mathbf{a}_n$ to $\mathbf{b}_n$) gradually aligns with the direction $\overline{\mathbf{u}}$. The assumption says that in this situation the dissimilarity $D\mathbf{a}_n\mathbf{b}_n$ and the Euclidean distance $|\mathbf{b}_n - \mathbf{a}_n|$ are comeasurable in the small: neither of them tends to zero infinitely faster than the other.*

We now generalize this function to apply to any vector $\mathbf{u}$, not just the unit one.

Definition 3.29. The function

$$F : \mathbb{T} \cup \{(\mathbf{x}, \mathbf{0}) : \mathbf{x} \in \mathfrak{S}\} \to \mathbb{R}$$

defined as

$$F(\mathbf{x}, \mathbf{u}) = \begin{cases} \lim_{s \to 0+} \frac{D\mathbf{x}[\mathbf{x}+\mathbf{u}s]}{s} & \text{if } \mathbf{u} \neq \mathbf{0} \\ 0 & \text{if } \mathbf{u} = \mathbf{0} \end{cases}, \tag{3.37}$$

is called a *submetric function*.

The standard term for $F(\mathbf{x}, \mathbf{u})$ in differential geometry is "metric function." It can, however, easily be confused with a metric on the space of stimuli, such as $G\mathbf{ab}$. To prevent this confusion, we use the nonstandard term "submetric function."

Theorem 3.30. *$F(\mathbf{x}, \mathbf{u})$ is well-defined for any $(\mathbf{x}, \mathbf{u}) \in \mathbb{T} \cup \{(\mathbf{x}, \mathbf{0}) : \mathbf{x} \in \mathfrak{S}\}$. It is positive for $\mathbf{u} \neq \mathbf{0}$, continuous in $(\mathbf{x}, \mathbf{u})$, and Euler homogeneous in $\mathbf{u}$.*

Euler homogeneity in $\mathbf{u}$ means that for any $k > 0$, $F(\mathbf{x}, k\mathbf{u}) = kF(\mathbf{x}, \mathbf{u})$. See the Appendix to this chapter for a proof.

Assumption $\mathcal{E}2$ can now be strengthened as follows.

Theorem 3.31. *For any $\mathbf{a}_n, \mathbf{b}_n \in \mathfrak{s} \subset \mathfrak{S}$, if $\mathfrak{s}$ is compact and $\mathbf{a}_n \leftrightarrow \mathbf{b}_n$ ($\mathbf{a}_n \neq \mathbf{b}_n$), then*

$$\frac{D\mathbf{a}_n\mathbf{b}_n}{F(\mathbf{a}_n, \mathbf{b}_n - \mathbf{a}_n)} \to 1.$$

Indeed, rewrite

$$\frac{D\mathbf{a}_n\mathbf{b}_n}{F(\mathbf{a}_n, \mathbf{b}_n - \mathbf{a}_n)} = \frac{D\mathbf{a}_n\mathbf{b}_n}{F\left(\mathbf{a}_n, \overline{\mathbf{b}_n - \mathbf{a}_n}\right)|\mathbf{b}_n - \mathbf{a}_n|},$$

and denote either lim inf or lim sup of this ratio by l. There is an infinite subsequence of $(\mathbf{a}_n, \mathbf{b}_n)$ (without loss of generality, the sequence itself) for which

$$\frac{D\mathbf{a}_n\mathbf{b}_n}{F\left(\mathbf{a}_n, \overline{\mathbf{b}_n - \mathbf{a}_n}\right)|\mathbf{b}_n - \mathbf{a}_n|} \to l.$$

But within a compact set $\mathfrak{s}$ one can always select from this sequence $(\mathbf{a}_n, \mathbf{b}_n)$ a subsequence with $\mathbf{a}_n \leftrightarrow \mathbf{x}$, $\mathbf{b}_n \leftrightarrow \mathbf{x}$, for some $\mathbf{x}$; and due to the compactness of the set $\overline{\mathfrak{u}}$ of all unit directions, one can always select a subsequence of this subsequence with $\overline{\mathbf{b}_n - \mathbf{a}_n} \to \overline{\mathbf{u}}$, for some $\overline{\mathbf{u}}$. In this resulting subsequence (again, without changing the indexation for convenience):

$$F\left(\mathbf{a}_n, \overline{\mathbf{b}_n - \mathbf{a}_n}\right) \to F(\mathbf{a}, \overline{\mathbf{u}}),$$

whence

$$\frac{D\mathbf{a}_n\mathbf{b}_n}{|\mathbf{b}_n - \mathbf{a}_n|} \to lF(\mathbf{a}, \overline{\mathbf{u}}).$$

By Assumption $\mathcal{E}2$ then, $l = 1$. Since this result holds for both lim inf and lim sup of the original ratio, the statement of the theorem follows.

3.8.3 Indicatrices

Definition 3.32. The function

$$\mathbf{1} : \mathbb{T} \to \mathbb{U}^n$$

defined by

$$\mathbf{1}(\mathbf{a}, \mathbf{u}) = \frac{\mathbf{u}}{F(\mathbf{a}, \mathbf{u})}$$

is called the *radius-vector function* associated with (or corresponding to) the submetric function $F(\mathbf{a}, \mathbf{u})$. The values of this function are referred to as *radius-vectors*. For a fixed $\mathbf{a} \in \mathfrak{S}$, the function $\mathbf{u} \mapsto \mathbf{1}(\mathbf{a}, \mathbf{u})$ is called the *indicatrix centered at* (or *attached to*) the point $\mathbf{a}$. The set

$$\mathbb{I}_{\mathbf{a}} = \left\{ \mathbf{u} \in \mathbb{U}^n : F(\mathbf{a}, \mathbf{u}) \leq 1 \right\}$$

is called the *body* of this indicatrix, and the set

$$\delta\mathbb{I}_{\mathbf{a}} = \left\{ \mathbf{u} \in \mathbb{U}^n : F(\mathbf{a}, \mathbf{u}) = 1 \right\}$$

is called its *boundary*.

Figure 3.18 provides an illustration for the relationship between $F(\mathbf{a}, \mathbf{u})$ and $\mathbf{1}(\mathbf{a}, \mathbf{u})$.

Note that $\{\mathbf{a}\} \times \mathbb{I}_{\mathbf{a}}$ is a subset of the tangent space $\mathbb{T}_{\mathbf{a}}$. Note also that the body (or the boundary) of an indicatrix is a set of vectors in $\mathbb{U}^n$ emanating from a common origin. The boundary should not be thought of as the set of the endpoints of the radius-vectors: the latter set does not determine the indicatrix uniquely, as one should also know the position of the origin within the boundary (see Figure 3.19). Not all points within a given set of endpoints may serve as points of origin: by definition, there can be no endpoint A on the boundary which is not connected to the origin O by a vector $\overrightarrow{OA} \in \delta\mathbb{I}_{\mathbf{a}}$, and the boundary cannot have two codirectional but nonidentical vectors $\overrightarrow{OA}$ and $\overrightarrow{OB}$ (see Figure 3.20): indeed, if

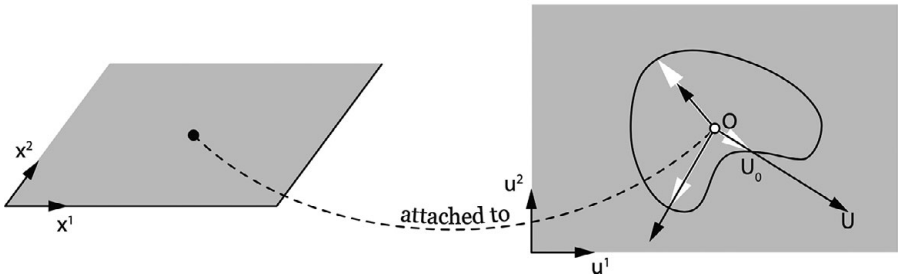


Figure 3.18 *An indicatrix (right) attached to a point in plane (left). The value of the submetric function F at this point and any vector $\overrightarrow{OU}$ is computed as the ratio of $\overrightarrow{OU}$ to the codirectional radius-vector of the indicatrix, $\overrightarrow{OU_0}$ (shown in white).*

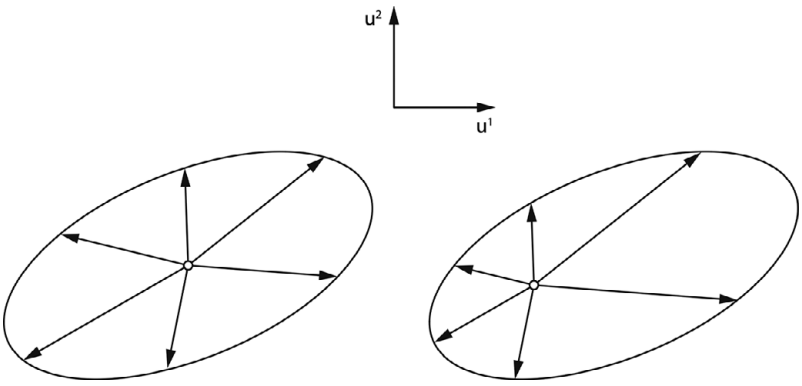


Figure 3.19 *The two indicatrices are different (consist of different vectors) although they have identical sets of endpoints.*

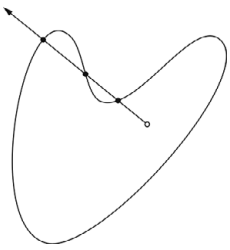


Figure 3.20 *This combination of a set of endpoints with a position of the origin does not form an indicatrix, because a radius-vector from the origin (shown by the open circle) intersects the boundary at more than one point.*

$$\frac{\overrightarrow{OA}}{\overrightarrow{OB}} = k \neq 1,$$

then

$$\frac{F\left(\mathbf{a}, \overrightarrow{OA}\right)}{F\left(\mathbf{a}, \overrightarrow{OB}\right)} = k,$$

so one of the vectors $\overrightarrow{OA}$ and $\overrightarrow{OB}$ does not belong to $\delta\mathbb{I}_{\mathbf{a}}$.

Figure 3.21 offers a geometric interpretation for measuring the length of a smooth path, to be rigorously justified later.

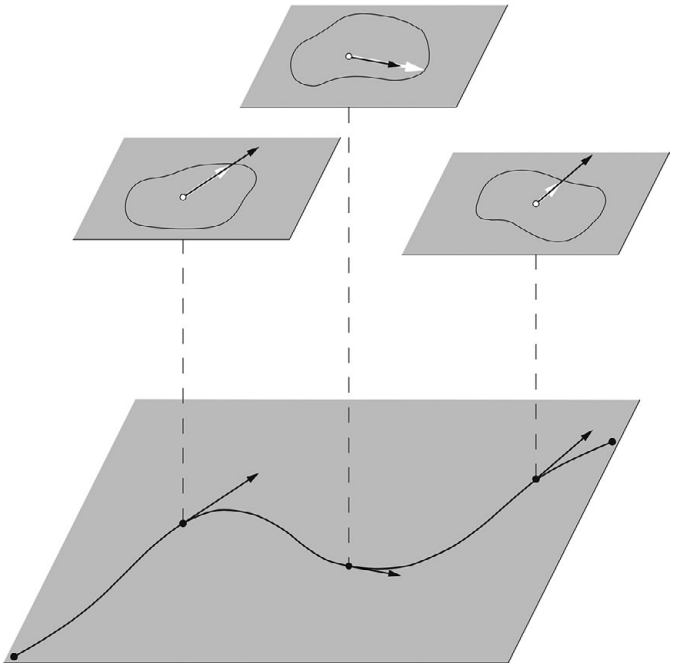


Figure 3.21 *Geometric interpretation of how indicatrices measure tangents to a smooth path: by centering the indicatrix $\mathbb{I}_{\mathbf{f}(x)}$ at each point $\mathbf{f}(x)$, one measures the magnitude of the tangent at this point by relating it to the codirectional radius-vector of the indicatrix, as explained in Figure 3.18. The length of the path $\mathbf{f}|[a,b]$ is then obtained by integrating this magnitude from a to b. For the conventional Euclidean length all indicatrices are unit-radius circles.*

We now list basic, almost obvious, properties of the unit vector function and the corresponding indicatrices.

Theorem 3.33. *The following statements hold true:*

(i) $\mathbf{1}(\mathbf{a},\mathbf{u})$ *is continuous;*

(ii) $\mathbf{1}(\mathbf{a},k\mathbf{u}) = \mathbf{1}(\mathbf{a},\mathbf{u})$ for all $(\mathbf{a},\mathbf{u}) \in \mathbb{T}$ and all $k > 0$ (Euler homogeneity in $\mathbf{u}$ of order zero);

(iii) for any $\mathbf{a} \in \mathfrak{S}$*, the mapping* $\overline{\mathbf{u}} \mapsto \mathbf{1}(\mathbf{a},\overline{\mathbf{u}})$ *is a homeomorphism;*

(iv) $\mathbb{I}_{\mathbf{a}}$ *is a compact set in* $\mathbb{U}^n$*;*

(v) $\delta\mathbb{I}_{\mathbf{a}}$ *is a compact set in* $\mathbb{U}^n$*;*

(vi) for any $\mathbf{a} \in \mathfrak{S}$*, there are two positive reals* $k_{\mathbf{a}}, K_{\mathbf{a}}$ *such that*

$$k_{\mathbf{a}} \leq |\mathbf{1}(\mathbf{a},\mathbf{u})| \leq K_{\mathbf{a}}$$

for all $\mathbf{u} \in \mathbb{U}$, and the values $k_{\mathbf{a}}, K_{\mathbf{a}}$ *are attained by* $\mathbf{1}(\mathbf{a},\mathbf{u})$ *at some* $\mathbf{u}$.

The proof of Propositions (i) and (ii) follows from the continuity and Euler homogeneity of $F(\mathbf{a},\mathbf{u})$. Denoting $\mathbf{1}(\mathbf{a},\overline{\mathbf{u}})$ by $\widetilde{\boldsymbol{u}}$, Proposition (iii) follows from the relations

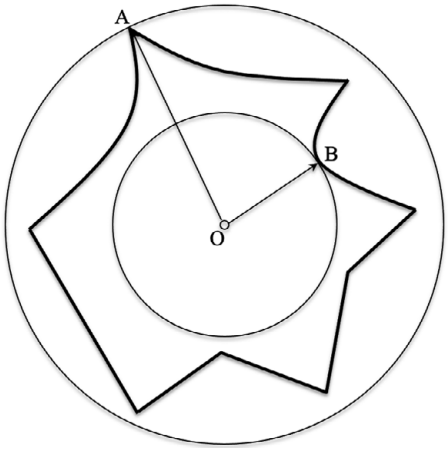


Figure 3.22 *A planar indicatrix (whose origin point O is attached to a point* $\mathbf{a}$ *in* $\mathfrak{S}$*) is sandwiched between two concentric circles of radii* $\left|\overrightarrow{OA}\right| = K_{\mathbf{a}}$ *and* $\left|\overrightarrow{OB}\right| = k_{\mathbf{a}}$.

$$\frac{\widetilde{\boldsymbol{u}}}{|\widetilde{\boldsymbol{u}}|} = \overline{\mathbf{u}}$$

and

$$\widetilde{\mathbf{u}} = \frac{\overline{\mathbf{u}}}{F(\mathbf{a}, \overline{\mathbf{u}})},$$

because both these functions are injective and continuous. The continuous function $\overline{\mathbf{u}} \mapsto \widetilde{\mathbf{u}}$ induces the continuous function $k\overline{\mathbf{u}} \mapsto k\widetilde{\mathbf{u}}$ for all $k \in [0,1]$, and (iv)–(v) then follow from the compactness of the unit Euclidean ball $\{k\overline{\mathbf{u}} : k \in [0,1]\}$ and the unit Euclidean sphere $\{\overline{\mathbf{u}}\}$. The continuous mapping $\overline{\mathbf{u}} \mapsto \mathbf{1}(\mathbf{a}, \overline{\mathbf{u}})$ of the compact unit Euclidean sphere should attain a maximum value $K_{\mathbf{a}}$ and a minimum value $k_{\mathbf{a}}$, and we get (vi) due to (ii).

Based on Theorem 3.33, we can think of an indicatrix boundary as a homeomorphically "deformed" Euclidean $(n-1)$-sphere "sandwiched" between two concentric Euclidean $(n-1)$-spheres of radii $k_{\mathbf{a}} > 0$ and $K_{\mathbf{a}} \geq k_{\mathbf{a}}$. Figure 3.22 illustrates this for $n = 2$.

3.8.4 Convex Combinations and Hulls

To further investigate the properties of indicatrices, we need to recall certain notions from linear algebra. In the vector space $\mathbb{U}^n$, a linear combination

$$\mathbf{u} = \lambda_1 \mathbf{v}_1 + \cdots + \lambda_m \mathbf{v}_m,\ m \geq 1 \tag{3.38}$$

is called a *convex combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ if $\lambda_i \geq 0$ for $i = 1, \ldots, m$, and

$$\lambda_1 + \cdots + \lambda_m = 1.$$

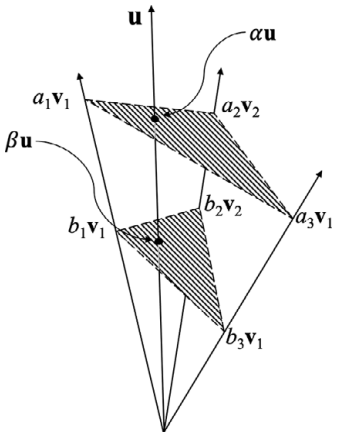


Figure 3.23 *Illustration for Lemma 3.34: a direction within the cone formed by* $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ *first crosses the lower facet and then the higher facet.*

From a geometric point of view, the set of convex combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ forms an $(m-1)$-dimensional facet with vertices $\mathbf{v}_1, \ldots, \mathbf{v}_m$. The following therefore is obviously true.

Lemma 3.34. *If* $\alpha\mathbf{u}$ *is a convex combination of* $a_1\mathbf{v}_1, \ldots, a_m\mathbf{v}_m$ *and* $\beta\mathbf{u}$ *is a convex combination of* $b_1\mathbf{v}_1, \ldots, b_m\mathbf{v}_m$*, with* $a_i \geq b_i$ *for* $i = 1, \ldots, m$ *and at least one inequality being strict, then* $\alpha > \beta$.

Figure 3.23 provides an illustration.

Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are called *affinely dependent* if, for some $\gamma_1, \ldots, \gamma_m$, not all zero:

$$\begin{array}{c} \gamma_1\mathbf{v}_1 + \cdots + \gamma_m\mathbf{v}_m = \mathbf{0} \\ \gamma_1 + \cdots + \gamma_m = 0 \end{array}. \tag{3.39}$$

If $\mathbf{u}$ is a convex combination of affinely dependent vectors, we have simultaneously

$$\left\{ \begin{array}{l} \lambda_1\mathbf{v}_1 + \cdots + \lambda_m\mathbf{v}_m = \mathbf{u} \\ \gamma_1\mathbf{v}_1 + \cdots + \gamma_m\mathbf{v}_m = \mathbf{0} \end{array} \right.,$$

where

$$\left\{ \begin{array}{l} \lambda_1 + \cdots + \lambda_m = 1 \\ \gamma_1 + \cdots + \gamma_m = 0 \end{array} \right.,$$

all λs are non-negative and some γs are nonzero (which means that at least one of them is positive and at least one negative). To exclude trivial cases, let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be pairwise distinct and let $\lambda_i > 0$ for $i = 1, \ldots, m$. Let c be the minimum $\left|\frac{\lambda_i}{\gamma_i}\right|$ among all negative ratios $\frac{\lambda_i}{\gamma_i}$. Then at least one of the coefficients in the representation

$$\mathbf{u} = (\lambda_1 + c\gamma_1)\,\mathbf{v}_1 + \cdots + (\lambda_m + c\gamma_m)\,\mathbf{v}_m$$

is zero, while all other coefficients are non-negative and sum to 1. This means that $\mathbf{u}$ is a convex combination of at most $m-1$ elements of $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$, and we have

Lemma 3.35. *If $\mathbf{u} \in \mathbb{U}^n$ is a convex combination of affinely dependent $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{U}^n$, then $\mathbf{u}$ is a convex combination of some $m' < m$ elements of $\mathbf{v}_1, \ldots, \mathbf{v}_m$.*

The following corollary of the lemma is known as a Carathéodory theorem.

Corollary 3.36. *If $\mathbf{u}, \mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{U}^n$, $m > n+1$, and $\mathbf{u}$ is a convex combination of $\mathbf{v}_1, \ldots, \mathbf{v}_m$, then $\mathbf{u}$ is a convex combination of at most $n+1$ elements of $\mathbf{v}_1, \ldots, \mathbf{v}_m$.*

This follows from the fact that if $m > n+1$, any $\mathbf{v}_1, \ldots, \mathbf{v}_m$ in $\mathbb{U}^n$ are affinely dependent. Indeed, since $\operatorname{rank}(\mathbf{v}_1, \ldots, \mathbf{v}_m) \leq n$, there should exist reals $\alpha_1, \ldots, \alpha_m$, not all zero, such that the system of $n+1$ linear equations

$$\begin{cases} \alpha_1 \mathbf{v}_1 + \cdots + \alpha_m \mathbf{v}_m = \mathbf{0} \\ \alpha_1 + \cdots + \alpha_m = 0 \end{cases}$$

is satisfied.

A subset $\mathbb{V}$ of $\mathbb{U}^n$ is said to be *convex* if it contains any convex combination

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{y}, \quad \lambda \in [0,1],$$

of any two of its elements $\mathbf{x}, \mathbf{y}$. By induction from 2 to $(n+1)$-element subsets of $\mathbb{V}$ (which is sufficient by Corollary 3.36), we see that a convex set $\mathfrak{X} \subset \mathbb{U}^n$ contains all convex combinations of *all* finite subsets of $\mathbb{V}$.

For any $\mathfrak{X} \subset \mathbb{U}^n$ the set of all convex combinations of all $(n+1)$-tuples of elements of $\mathbb{V}$ is called the *convex hull* of $\mathbb{V}$ and is denoted $\operatorname{conv}\mathbb{V}$. Again, $\operatorname{conv}\mathbb{V}$ is, clearly, the set of all convex combinations of all finite subsets of $\mathbb{V}$, and it is the smallest convex subset of $\mathbb{U}^n$ containing $\mathbb{V}$.

Consider now an indicatrix $\mathbb{I}_{\mathbf{a}}$ and its convex hull. The following is obvious.

Lemma 3.37. *For any indicatrix $\mathbb{I}_{\mathbf{a}}$, $\operatorname{conv}\mathbb{I}_{\mathbf{a}}$ is compact in $\mathbb{U}^n$.*

Let now $\mathbf{u} \in \operatorname{conv}\mathbb{I}_{\mathbf{a}}$. Then, for some $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{I}_{\mathbf{a}}$ and some non-negative reals $\lambda_1, \ldots, \lambda_m$ that sum to 1,

$$\mathbf{u} = \lambda_1 \mathbf{v}_1 + \cdots + \lambda_m \mathbf{v}_m.$$

But then

$$\begin{aligned} |\mathbf{u}| &= |\lambda_1 \mathbf{v}_1 + \cdots + \lambda_m \mathbf{v}_m| \leq \lambda_1 |\mathbf{v}_1| + \cdots + \lambda_m |\mathbf{v}_m| \\ &\leq (\lambda_1 + \cdots + \lambda) K_{\mathbf{a}} = K_{\mathbf{a}}, \end{aligned}$$

where $K_{\mathbf{a}}$ denotes $\max_{\mathbf{u} \in \mathbb{I}_{\mathbf{a}}} |\mathbf{u}|$ (whose existence is stated in Theorem 3.33(v)). We have therefore

Lemma 3.38. *For any $\mathbf{a} \in \mathfrak{S}$,*

$$\max_{\mathbf{u} \in \operatorname{conv}\mathbb{I}_{\mathbf{a}}} |\mathbf{u}| = \max_{\mathbf{u} \in \mathbb{I}_{\mathbf{a}}} |\mathbf{u}|.$$

Definition 3.39. For any $(\mathbf{a},\mathbf{u}) \in \mathbb{T}$, the quantity

$$\kappa\,(\mathbf{a},\mathbf{u}) = \max\{\alpha > 0 : \alpha\mathbf{1}(\mathbf{a},\mathbf{u}) \in \mathrm{conv}\mathbb{I}_{\mathbf{a}}\}$$

is called the *maximal production factor* for $\mathbf{u}$ in $\mathbb{I}_{\mathbf{a}}$, and the vector $\kappa\,(\mathbf{a},\mathbf{u})\,\mathbf{1}(\mathbf{a},\mathbf{u})$ is called the *maximal production* of (or *maximally produced*) $\mathbf{u}$ in $\mathbb{I}_{\mathbf{a}}$.

This is clearly a well-defined function, because it follows from the compactness of $\mathrm{conv}\mathbb{I}_{\mathbf{a}}$ that

Lemma 3.40. *For any $\mathbf{a} \in \mathfrak{S}$, every $\mathbf{u} \in \mathbb{U}^n$ has its maximal production in $\mathbb{I}_{\mathbf{a}}$.*

The following statement holds because $\alpha\mathbf{u}$ and $\mathbf{u}$ have one and the same maximal production in $\mathbb{I}_{\mathbf{a}}$.

Lemma 3.41. *The function $\kappa\,(\mathbf{a},\mathbf{u})$ is Euler homogeneous of zero order:*

$$\kappa\,(\mathbf{a},\alpha\mathbf{u}) = \kappa\,(\mathbf{a},\mathbf{u})\,.$$

Finally, we need to observe the following.

Lemma 3.42. *For any $(\mathbf{a},\mathbf{u}) \in \mathbb{T}$, the maximal production* of $\mathbf{u}$ in $\mathbb{I}_{\mathbf{a}}$ *can be presented as a convex combination of n (not necessarily distinct) radius-vectors $\mathbf{v}_1,\ldots,\mathbf{v}_n \in \delta\mathbb{I}_{\mathbf{a}}$.*

See the Appendix for a proof.

Figure 3.24 provides an illustration for this lemma on three-dimensional indicatrices. (It also illustrates the useful notion of the degree of flatness for a radius vector within the body of the indicatrix.)

3.8.5 Minimal Submetric Function and Convex Hulls of Indicatrices

In this section we consider the problem of finding a *geodesic in the small*, a shortest path connecting stimuli $\mathbf{a}$ and $\mathbf{a}+\mathbf{u}s$ as $s \to 0$. It will be established later (Section 3.8.6) that $G\mathbf{a}\,(\mathbf{a}+\mathbf{u}s)$ in $\mathfrak{S} \subseteq \mathbb{R}^n$ can be approximated by concatenation of $m \leq n$ straight-line segments with lengths $F\,(\mathbf{a},\mathbf{u}_i)\,s$ for some vectors $\mathbf{u}_1,\ldots,\mathbf{u}_m$ summing to $\mathbf{u}$. So we begin with investigating the minimal value for certain sums of $F\,(\mathbf{a},\mathbf{u}_i)$.

Definition 3.43. A sequence of vectors $(\mathbf{u}_1,\ldots,\mathbf{u}_m)$ in $\mathbb{U}^n$, $m \geq 1$, is said to form a *minimizing vector chain* for a line element $(\mathbf{a},\mathbf{u}) \in \mathbb{T}$ if

$$\mathbf{u} = \mathbf{u}_1 + \cdots + \mathbf{u}_m$$

and

$$F\,(\mathbf{a},\mathbf{u}_1) + \cdots + F\,(\mathbf{a},\mathbf{u}_m) = \min\{F\,(\mathbf{a},\mathbf{v}_1) + \cdots + F\,(\mathbf{a},\mathbf{v}_k)\}\,,$$

where the minimum is taken over all $k \geq 1$ and all finite sequences $(\mathbf{v}_1,\ldots,\mathbf{v}_k)$ in $\mathbb{U}^n$ such that

$$\mathbf{u} = \mathbf{v}_1 + \cdots + \mathbf{v}_k.$$

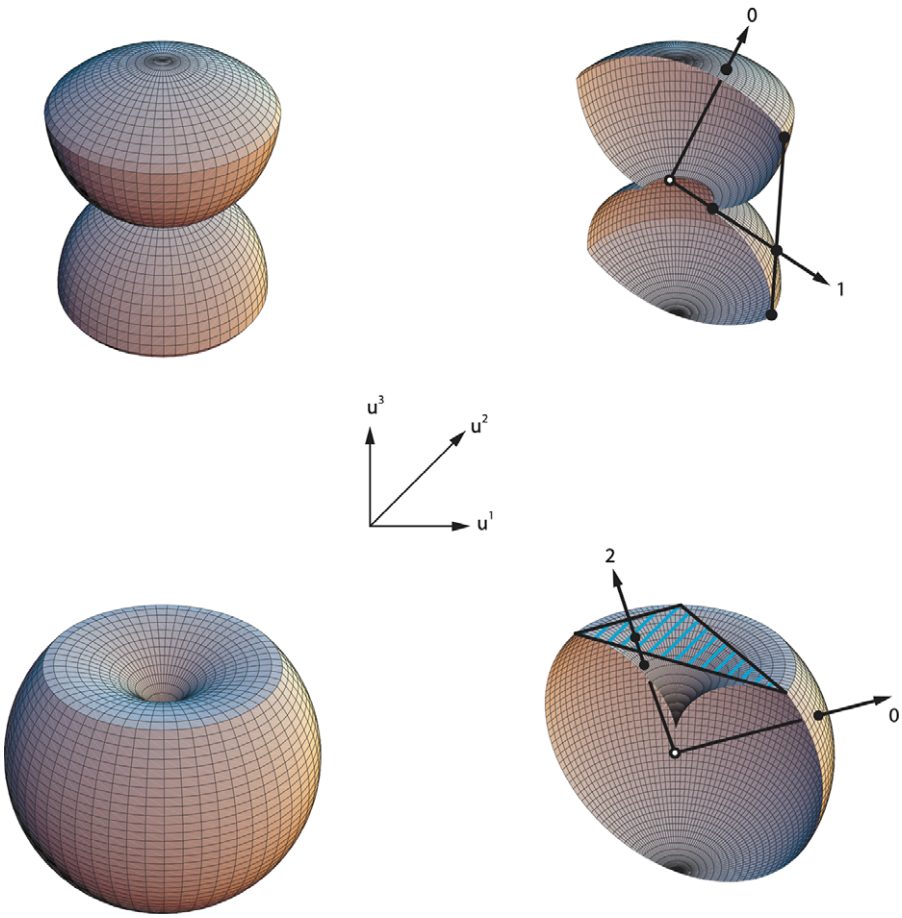


Figure 3.24 *Two indicatrices in* $\mathbb{U}^3$ *(left) and their cross-sections (right) showing the position of the origin (white dots). The maximal productions of two vectors are shown in each of the indicatrices, as parts of the vectors between the origin to the farthest black dot. The number attached to a vector* $\mathbf{v}$ *shows the degree of flatness* $r-1$ *of the indicatrix in the direction* $\mathbf{v}$*, where* r *is the maximum number of linearly independent radius-vectors whose convex combination equals the maximum production of* $\mathbf{v}$ *in the body of the indicatrix.*

Note that this definition does not require that $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be pairwise distinct, so a minimizing chain for $(\mathbf{a}, \mathbf{u})$ may, for example, be $\left\{\frac{1}{n}\mathbf{u}, \ldots, \frac{1}{n}\mathbf{u}\right\}$(which is equivalent to $\mathbf{u}$ alone being a minimizing vector chain for $(\mathbf{a}, \mathbf{u})$ too). Note also, that if $(\mathbf{u}_1, \ldots, \mathbf{u}_m)$ is a minimizing chain, then so is any permutation thereof.

Theorem 3.44. *A minimizing chain for any* $(\mathbf{a}, \mathbf{u}) \in \mathbb{T}$ *exists and consists of* n *(not necessarily distinct) nonzero vectors* $\mathbf{u}_1, \ldots, \mathbf{u}_n$*, such that*

$$F(\mathbf{a}, \mathbf{u}_1) + \cdots + F(\mathbf{a}, \mathbf{u}_n) = \frac{F(\mathbf{a}, \mathbf{u})}{\kappa(\mathbf{a}, \mathbf{u})},$$

where $\kappa(\mathbf{a}, \mathbf{u})$ *is the maximal production factor for* $\mathbf{u}$ *in* $\mathbb{I}_{\mathbf{a}}$.

To prove this, we fix $\kappa\,(\mathbf{a},\mathbf{u}) = \kappa$ as we deal with a fixed $(\mathbf{a},\mathbf{u})$. Consider the maximal production $\kappa\mathbf{1}\,(\mathbf{a},\mathbf{u})$ of $\mathbf{u}$. By Lemma 3.42, it can be presented as a convex combination of some n radius-vectors $\widetilde{\mathbf{v}}_1,\ldots,\widetilde{\mathbf{v}}_n$ in $\delta\mathbb{I}_{\mathbf{a}}$:

$$\kappa\mathbf{1}\,(\mathbf{a},\mathbf{u}) = \lambda_1\widetilde{\mathbf{v}}_1 + \cdots + \lambda_n\widetilde{\mathbf{v}}_n,$$

where all coefficients are non-negative and sum to 1. Then, denoting

$$\mathbf{v}_i = \frac{\lambda_i}{\kappa}\widetilde{\mathbf{v}}_i, \qquad i = 1,\ldots,n,$$

we have

$$\mathbf{1}\,(\mathbf{a},\mathbf{u}) = \mathbf{v}_1 + \cdots + \mathbf{v}_n$$

and

$$F\,(\mathbf{a},\mathbf{v}_1) + \cdots + F\,(\mathbf{a},\mathbf{v}_n) = \frac{1}{\kappa}.$$

We prove now that for any $\mathbf{w}_1,\ldots,\mathbf{w}_m$ in $\mathbb{U}^n$, if

$$\mathbf{1}\,(\mathbf{a},\mathbf{u}) = \mathbf{w}_1 + \cdots + \mathbf{w}_m,$$

then

$$F\,(\mathbf{a},\mathbf{w}_1) + \cdots + F\,(\mathbf{a},\mathbf{w}_m) = \delta \geq \frac{1}{\kappa}.$$

Indeed, we have

$$\mathbf{1}\,(\mathbf{a},\mathbf{u}) = F\,(\mathbf{a},\mathbf{w}_1)\,\mathbf{1}\,(\mathbf{a},\mathbf{w}_1) + \cdots + F\,(\mathbf{a},\mathbf{w}_m)\,\mathbf{1}\,(\mathbf{a},\mathbf{w}_m)$$

and

$$\frac{1}{\delta}\mathbf{1}\,(\mathbf{a},\mathbf{u}) = \frac{F\,(\mathbf{a},\mathbf{w}_1)}{\delta}\mathbf{1}\,(\mathbf{a},\mathbf{w}_1) + \cdots + \frac{F\,(\mathbf{a},\mathbf{w}_{n+1})}{\delta}\mathbf{1}\,(\mathbf{a},\mathbf{w}_m)\,.$$

That is, $\frac{1}{\delta}\mathbf{1}\,(\mathbf{a},\mathbf{u})$ is a convex combination of m radius-vectors of $\delta\mathbb{I}_{\mathbf{a}}$. But then

$$\frac{1}{\delta} \leq \kappa.$$

It follows that $(\mathbf{v}_1,\ldots,\mathbf{v}_n)$ is a minimizing vector chain for $(\mathbf{a},\mathbf{1}\,(\mathbf{a},\mathbf{u}))$, with

$$F\,(\mathbf{a},\mathbf{v}_1) + \cdots + F\,(\mathbf{a},\mathbf{v}_m) = \frac{1}{\kappa}.$$

The statement of the theorem obtains by putting $\mathbf{u}_i = F\,(\mathbf{a},\mathbf{u})\,\mathbf{v}_i$, $i = 1,\ldots,n$.

We introduce now one of the central notions of the theory.

Definition 3.45. For any $(\mathbf{a},\mathbf{u}) \in \mathbb{T}\cup\{(\mathbf{x},\mathbf{0}) : \mathbf{x} \in \mathfrak{S}\}$, the function

$$\widehat{F}\,(\mathbf{a},\mathbf{u}) = \begin{cases} \frac{F(\mathbf{a},\mathbf{u})}{\kappa(\mathbf{a},\mathbf{u})} & \text{if } \mathbf{u} \neq \mathbf{0} \\ 0 & \text{if } \mathbf{u} = \mathbf{0} \end{cases}$$

is called the *minimal submetric function*.

Clearly

$$\widehat{F}(\mathbf{a},\mathbf{u}) \leq F(\mathbf{a},\mathbf{u}).$$

Theorem 3.46. *The minimal submetric function* $\widehat{F}(\mathbf{a},\mathbf{u})$ *has all the properties of a submetric function: it is positive for* $\mathbf{u} \neq \mathbf{0}$*, Euler homogeneous, and continuous.*

See the Appendix for a proof.

Theorem 3.47. *The indicatrix at* $\mathbf{a} \in \mathfrak{S}$ *associated with* $\widehat{F}(\mathbf{a},\mathbf{u})$*,*

$$\mathbf{u} \mapsto \widehat{\mathbf{1}}(\mathbf{a},\mathbf{u}) = \frac{\mathbf{u}}{\widehat{F}(\mathbf{a},\mathbf{u})},$$

has the body

$$\widehat{\mathbb{I}}_{\mathbf{a}} = \left\{\mathbf{u} \in \mathbb{U}^n : \widehat{F}(\mathbf{a},\mathbf{u}) \leq 1\right\} = \text{conv}\mathbb{I}_{\mathbf{a}},$$

where $\mathbb{I}_{\mathbf{a}}$ *is the body of the indicatrix* $\mathbf{u} \mapsto \mathbf{1}(\mathbf{a},\mathbf{u})$ *associated with* $F(\mathbf{a},\mathbf{u})$*. The boundary*

$$\widehat{\delta\mathbb{I}}_{\mathbf{a}} = \left\{\mathbf{u} \in \mathbb{U}^n : \widehat{F}(\mathbf{a},\mathbf{u}) = 1\right\}$$

of the indicatrix $\mathbf{u} \mapsto \widehat{\mathbf{1}}(\mathbf{a},\mathbf{u})$ *is the set of all maximally produced radius-vectors of the indicatrix* $\mathbf{u} \mapsto \mathbf{1}(\mathbf{a},\mathbf{u})$*.*

This is essentially a summary of the results established so far. To prove the second statement of the theorem, by Lemma 3.40 and Theorem 3.44, the maximal production $\kappa(\mathbf{a},\mathbf{u})\mathbf{1}(\mathbf{a},\mathbf{u})$ of $\mathbf{u}$ in $\mathbb{I}_{\mathbf{a}}$ exists for every $\mathbf{u}$, and

$$\widehat{F}(\mathbf{a},\mathbf{1}(\mathbf{a},\mathbf{u})) = \frac{1}{\kappa(\mathbf{a},\mathbf{u})}.$$

It follows that $\widehat{F}(\mathbf{a},\mathbf{u}) = 1$ if and only if

$$\mathbf{u} = \kappa(\mathbf{a},\mathbf{u})\mathbf{1}(\mathbf{a},\mathbf{u}).$$

To prove the first statement of the theorem, by Lemma 3.42, $\kappa(\mathbf{a},\mathbf{u})\mathbf{1}(\mathbf{a},\mathbf{u})$ is a convex combination of some vectors $\mathbf{v}_1,\ldots,\mathbf{v}_n$ in $\mathbb{I}_{\mathbf{a}}$. But then $c\kappa(\mathbf{a},\mathbf{u})\mathbf{1}(\mathbf{a},\mathbf{u})$ is a convex combination of $c\mathbf{v}_1,\ldots,c\mathbf{v}_n \in \mathbb{I}_{\mathbf{a}}$ for any $c \in [0,1]$. It is clear then that $\text{conv}\mathbb{I}_{\mathbf{a}}$ consists of all vectors

$$\mathbf{u} = c\kappa(\mathbf{a},\mathbf{u})\mathbf{1}(\mathbf{a},\mathbf{u}), \qquad c \in [0,1].$$

But these are precisely the vectors satisfying $\widehat{F}(\mathbf{a},\mathbf{u}) \leq 1$. This completes the proof.

It follows from this theorem that $\widehat{\mathbf{1}}(\mathbf{a},\mathbf{u})$, $\widehat{\mathbb{I}}_{\mathbf{a}}$, and $\widehat{\delta\mathbb{I}}_{\mathbf{a}}$ have all the properties listed in Theorem 3.33. If $\delta\mathbb{I}_{\mathbf{a}}$ is a homeomorphically deformed Euclidean sphere sandwiched between two Euclidean spheres of radii $k_{\mathbf{a}}$ and $K_{\mathbf{a}}$, then $\widehat{\delta\mathbb{I}}_{\mathbf{a}}$ is a homeomorphically deformed (but convex) Euclidean sphere sandwiched between two Euclidean spheres of radii $k^*_{\mathbf{a}}$ and $K_{\mathbf{a}}$ (where $k^*_{\mathbf{a}} \geq k_{\mathbf{a}}$ and $K_{\mathbf{a}}$ is the same for $\delta\mathbb{I}_{\mathbf{a}}$ and $\widehat{\delta\mathbb{I}}_{\mathbf{a}}$, as stated in Lemma 3.38). Figure 3.25 illustrates this using the indicatrix shown in Figure 3.22. Figure 3.26 shows the convex hulls of the indicatrices shown in Figure 3.24.

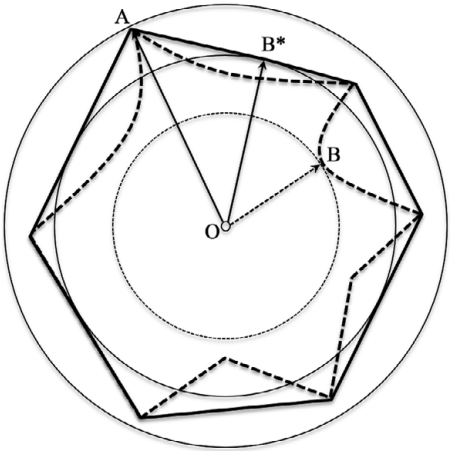


Figure 3.25 *The convex hull of the indicatrix body shown in Figure 3.22 is sandwiched between* $\left|\overrightarrow{OA}\right| = K_{\mathbf{a}}$ *(the same as for the indicatrix itself) and* $\left|\overrightarrow{OB^*}\right| = k^*_{\mathbf{a}}$ *which is greater than* $\left|\overrightarrow{OB}\right| = k_{\mathbf{a}}$.

3.8.6 Length and Metric in Euclidean Spaces

Definition 3.48. A submetric function $F(\mathbf{a},\mathbf{u})$ is called *convex* if for any $\mathbf{a} \in \mathfrak{S}$ and $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{U}^n$,

$$F(\mathbf{a},\mathbf{u}_1+\mathbf{u}_2) \leq F(\mathbf{a},\mathbf{u}_1) + F(\mathbf{a},\mathbf{u}_2).$$

Assume, excluding the trivial case, that $\mathbf{u}_1, \mathbf{u}_2$ are not both zero. If $\mathbb{I}_{\mathbf{a}}$ is convex, then the vector

$$\frac{F(\mathbf{a},\mathbf{u}_1)}{F(\mathbf{a},\mathbf{u}_1)+F(\mathbf{a},\mathbf{u}_2)}\mathbf{1}(\mathbf{a},\mathbf{u}_1) + \frac{F(\mathbf{a},\mathbf{u}_2)}{F(\mathbf{a},\mathbf{u}_1)+F(\mathbf{a},\mathbf{u}_2)}\mathbf{1}(\mathbf{a},\mathbf{u}_2) \in \mathbb{I}_{\mathbf{a}}. \tag{3.40}$$

This is equivalent to

$$F\left(\mathbf{a}, \frac{F(\mathbf{a},\mathbf{u}_1)}{F(\mathbf{a},\mathbf{u}_1)+F(\mathbf{a},\mathbf{u}_2)}\mathbf{1}(\mathbf{a},\mathbf{u}_1) + \frac{F(\mathbf{a},\mathbf{u}_2)}{F(\mathbf{a},\mathbf{u}_1)+F(\mathbf{a},\mathbf{u}_2)}\mathbf{1}(\mathbf{a},\mathbf{u}_2)\right) \leq 1.$$

But the left-hand side expression equals

$$\frac{F(\mathbf{a},\mathbf{u}_1+\mathbf{u}_2)}{F(\mathbf{a},\mathbf{u}_1)+F(\mathbf{a},\mathbf{u}_2)},$$

whence we see that $F(\mathbf{a},\mathbf{u})$ is convex. Conversely, if the expression above is ≤ 1, then (3.40) holds. Since it holds for any $\mathbf{u}_1, \mathbf{u}_2$, it also holds for $\lambda\mathbf{u}_1$, $(1-\lambda)\mathbf{u}_2$ for $0 \leq \lambda \leq 1$. But, as λ changes from 0 to 1, the expression

$$\frac{F(\mathbf{a},\lambda\mathbf{u}_1)}{F(\mathbf{a},\lambda\mathbf{u}_1)+F(\mathbf{a},\mathbf{u}_2)} = \frac{\lambda F(\mathbf{a},\mathbf{u}_1)}{\lambda F(\mathbf{a},\lambda\mathbf{u}_1)+(1-\lambda)F(\mathbf{a},(\mathbf{1}-\lambda)\mathbf{u}_2)}$$

runs through all values from 0 to 1 too. Since

$$\mathbf{1}(\mathbf{a},\lambda\mathbf{u}_1) = \mathbf{1}(\mathbf{a},\mathbf{u}_1), \mathbf{1}(\mathbf{a},(1-\lambda)\mathbf{u}_2) = \mathbf{1}(\mathbf{a},\mathbf{u}_2),$$

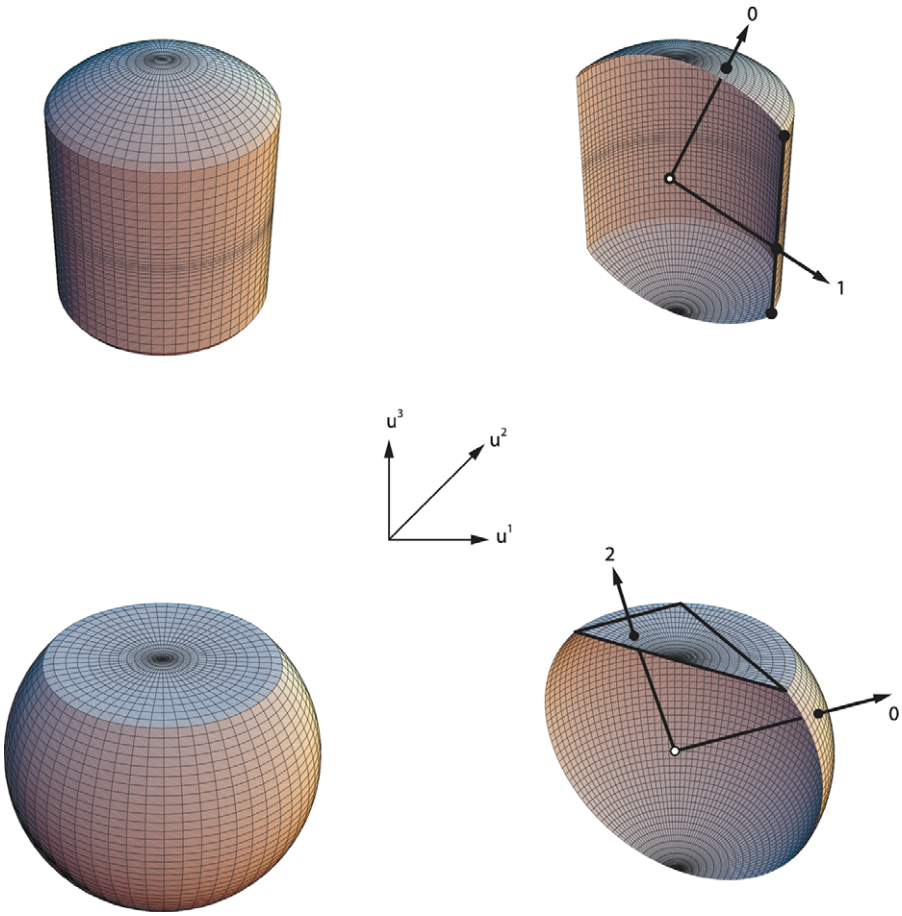


Figure 3.26 *The convex hulls of the indicatrices shown in Figure 3.24. The degree of flatness of codirectional radius-vectors remains unchanged.*

we have

$$\theta \mathbf{1}(\mathbf{a}, \mathbf{u}_1) + (1-\theta)\mathbf{1}(\mathbf{a}, \mathbf{u}_2) \in \mathbb{I}_{\mathbf{a}},$$

for any $0 \leq \theta \leq 1$. This means that $\mathbb{I}_{\mathbf{a}}$ is convex, and we have proved

Theorem 3.49. *$F(\mathbf{a}, \mathbf{u})$ is convex if and only if the body of the associated indicatrix $\mathbb{I}_{\mathbf{a}}$ at any point $\mathbf{a}$ is convex.*

From this and Theorem 3.47 we immediately have

Corollary 3.50. *For every submetric function F,*
(i) the corresponding minimal submetric function $\widehat{F}$ is convex;
(ii) $F \equiv \widehat{F}$ if and only if F is convex.

We also have

Corollary 3.51. *If a submetric function F is convex, then $\{\mathbf{u}\}$ is a minimizing vector chain for any line element $(\mathbf{a},\mathbf{u}) \in \mathbb{T}$.*

This follows from $F(\mathbf{a},\mathbf{u}) = \widehat{F}(\mathbf{a},\mathbf{u})$.

Of course, if F is convex, the following are also minimizing vector chains for $\mathbf{u} \in \mathbb{U}^n$: $\left\{\frac{1}{2}\mathbf{u}, \frac{1}{2}\mathbf{u}\right\}$, $\left\{\frac{1}{3}\mathbf{u}, \frac{2}{3}\mathbf{u}\right\}$,$\left\{\frac{1}{n}\mathbf{u}, \ldots, \frac{1}{n}\mathbf{u}\right\}$, etc. Moreover, if F is not strictly convex (i.e., the inequality in Definition 3.48 may be equality for some $\mathbf{u}_1, \mathbf{u}_2$), there may very well be minimizing chains involving vectors that are not collinear with $\mathbf{u}$.

We have now arrived at one of the central theorems in the theory.

Theorem 3.52. *The distance $G(\mathbf{x}, \mathbf{x}+\mathbf{u}s)$ is differentiable at $s = 0+$ for any $(\mathbf{x},\mathbf{u}) \in \mathbb{T}$, and*

$$\left.\frac{\mathrm{d}G(\mathbf{x},\mathbf{x}+\mathbf{u}s)}{\mathrm{d}s+}\right|_{s=0} = \lim_{s\to 0+}\frac{G(\mathbf{x},\mathbf{x}+\mathbf{u}s)}{s} = \widehat{F}(\mathbf{x},\mathbf{u}).$$

See the Appendix for a proof.

An important corollary to this theorem is as follows. Let $\mathbf{f}|[a,b]$ be a continuously differentiable path. Consider

$$\frac{G\mathbf{f}(t)\,\mathbf{f}(\tau)}{\widehat{F}(\mathbf{f}(t),\mathbf{f}(\tau)-\mathbf{f}(t))}, t<\tau.$$

By presenting it as

$$\frac{G\mathbf{f}(t)\left(\mathbf{f}(t)+\frac{\mathbf{f}(\tau)-\mathbf{f}(t)}{\tau-t}(\tau-t)\right)}{\widehat{F}\left(\mathbf{f}(t),\frac{\mathbf{f}(\tau)-\mathbf{x}(t)}{\tau-t}\mathbf{f}(\tau-t)\right)} = \frac{G\mathbf{f}(t)\left(\mathbf{f}(t)+\dot{\mathbf{f}}(\theta)(\tau-t)\right)}{\widehat{F}\left(\mathbf{f}(t),\dot{\mathbf{f}}(\theta)\,\mathbf{f}(\tau-t)\right)},$$

with $t \le \theta \le \tau$, we see that if $\tau - t \to 0+$ on $[a,b]$, the ratio tends to 1 (by Theorem 3.52 and because all functions involved are uniformly continuous on $[a,b]$). This establishes

Corollary 3.53. *For any smooth path $\mathbf{f}|[a,b]$ and $[t,\tau] \subset [a,b]$,*

$$\lim_{\tau-t\to 0+}\frac{G\mathbf{f}(t)\,\mathbf{f}(\tau)}{\widehat{F}(\mathbf{f}(t),\mathbf{f}(\tau)-\mathbf{f}(t))} = 1.$$

We are now ready to formulate the standard differential-geometric computation of the length of a continuously differentiable path by integration of the submetric function applied to its points and tangents.

Theorem 3.54. *For any continuously differentiable path $\mathbf{f}|[a,b]$,*

$$D\mathbf{f}|[a,b] = \int_a^b \widehat{F}\left(\mathbf{f}(t),\dot{\mathbf{f}}(t)\right)dt.$$

Indeed, by definition,

$$D\mathbf{f}|[a,b] = \lim_{\delta\mu\to 0}\sum G\mathbf{f}(t_i)\,\mathbf{f}(t_{i+1})$$

across all nets $\mu = \{\ldots, t_i, t_{i+1} \ldots\}$ partitioning $[a,b]$. This limit can be presented as

$$\lim_{\delta\mu\to 0} \sum \widehat{F}\left(\mathbf{f}\left(t_i\right), \mathbf{f}\left(t_{i+1}\right) - \mathbf{f}\left(t_i\right)\right) \frac{G\mathbf{f}\left(t_i\right)\mathbf{f}\left(t_{i+1}\right)}{\widehat{F}\left(\mathbf{f}\left(t_i\right), \mathbf{f}\left(t_{i+1}\right) - \mathbf{f}\left(t_i\right)\right)}.$$

By Corollary 3.53,

$$\lim_{\delta\mu\to 0} \frac{G\mathbf{f}\left(t_i\right)\mathbf{f}\left(t_{i+1}\right)}{\widehat{F}\left(\mathbf{f}\left(t_i\right), \mathbf{f}\left(t_{i+1}\right) - \mathbf{f}\left(t_i\right)\right)} = 1.$$

Then

$$\begin{aligned} D\mathbf{f}|\left[a,b\right] &= \lim_{\delta\mu\to 0} \sum \widehat{F}\left(\mathbf{f}\left(t_i\right), \mathbf{f}\left(t_{i+1}\right) - \mathbf{f}\left(t_i\right)\right) \\ &= \lim_{\delta\mu\to 0} \sum \widehat{F}\left(\mathbf{f}\left(t_i\right), \frac{\mathbf{f}\left(t_{i+1}\right) - \mathbf{f}\left(t_i\right)}{t_{i+1} - t_i}\right)\left(t_{i+1} - t_i\right). \end{aligned}$$

But

$$\lim_{\delta\mu\to 0} \widehat{F}\left(\mathbf{f}\left(t_i\right), \frac{\mathbf{f}\left(t_{i+1}\right) - \mathbf{f}\left(t_i\right)}{t_{i+1} - t_i}\right) = \widehat{F}\left(\mathbf{f}\left(t\right), \dot{\mathbf{f}}\left(t\right)\right)$$

and $\widehat{F}\left(\mathbf{f}\left(t\right), \dot{\mathbf{f}}\left(t\right)\right)$ is uniformly continuous on $[a,b]$. Hence

$$D\mathbf{f}|\left[a,b\right] = \lim_{\delta\mu\to 0} \sum \widehat{F}\left(\mathbf{f}\left(t_i\right), \dot{\mathbf{f}}\left(t_i\right)\right)\left(t_{i+1} - t_i\right) = \int_a^b \widehat{F}\left(\mathbf{f}\left(t\right), \dot{\mathbf{f}}\left(t\right)\right) dt,$$

completing the proof.

Since

$$\lim_{\tau - t\to 0+} \frac{\int_t^\tau \widehat{F}\left(\mathbf{f}\left(x\right), \dot{\mathbf{f}}\left(x\right)\right) dx}{\widehat{F}\left(\mathbf{f}\left(t\right), \frac{\mathbf{f}(\tau) - \mathbf{f}(t)}{\tau - t}\right)\left(\tau - t\right)} = 1,$$

we also have

Corollary 3.55. *For any continuously differentiable path* $\mathbf{f}|\left[a,b\right]$*, and* $[t,\tau] \subset [a,b]$,

$$\lim_{\tau - t\to 0+} \frac{G\mathbf{f}\left(t\right)\mathbf{f}\left(\tau\right)}{D\mathbf{f}|\left[t,\tau\right]} = 1.$$

3.8.7 Continuously Differentiable Paths and Intrinsic Metric G

Before proceeding, we need an auxiliary observation. The space $(\mathfrak{S}, E)$ being open, each point $\mathbf{p}$ in $\mathfrak{S}$ can be enclosed in a compact Euclidean ball

$$\mathfrak{B}\left(\mathbf{p}, r\right) = \left\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{p}| \leq r\right\} \subseteq \mathfrak{S},$$

and we can associate with any $\mathbf{p}$ the ball $\mathfrak{B}\left(\mathbf{p}, r\right)$ with the supremal value of $r_{\text{sup}}\left(\mathbf{p}\right)$ (including ∞). The observation is that, given any compact subset $\mathfrak{s}$ of $\mathfrak{S}$,

$$\inf_{\mathbf{p}\in\mathfrak{s}} r_{\text{sup}}\left(\mathbf{p}\right) = \min_{\mathbf{p}\in\mathfrak{s}} r_{\text{sup}}\left(\mathbf{p}\right) > 0.$$

A *straight-line segment* is defined as

$$\mathbf{s}(x) = \mathbf{a} + \mathbf{u}x, \qquad x \in [a,b],\ (\mathbf{a},\mathbf{u}) \in \mathbb{T}.$$

If $\mathbf{x}$ and $\mathbf{y}$ are within any ball $\mathfrak{B}(\mathbf{p},r)$, they can be connected by the straight-line segment

$$\mathbf{s}(x) = \mathbf{x} + \frac{\mathbf{y}-\mathbf{x}}{b-a}(x-a), \qquad x \in [a,b].$$

Concatenations of straight-line segments form piecewise linear paths, about which we have the following result.

Theorem 3.56. *For every path* $\mathbf{h}|[a,b]$ *connecting* $\mathbf{a}$ *to* $\mathbf{b}$ *one can find a piecewise linear path from* $\mathbf{a}$ *to* $\mathbf{b}$ *which is arbitrarily close to* $\mathbf{h}|[a,b]$ *pointwise and in its length.*

See the Appendix for a proof.

The straight-line segments are not indispensable in such an approximation. In fact, we can use the following "corner-rounding" procedure to replace any piecewise linear path with a continuously differentiable path. It is illustrated in Figure 3.27.

Let two adjacent straight-line segments be presented as

$$\mathbf{p}(t) = \begin{cases} \mathbf{a} + \overline{\mathbf{u}}_1 t & \text{if } t \in [-a,0] \\ \mathbf{a} + \overline{\mathbf{u}}_2 t & \text{if } t \in [0,b] \end{cases},$$

with $a,b > 0$. On a small interval $[-s,s]$,

$$D\mathbf{p}|[-s,s] = \int_{-s}^{0} \widehat{F}(\mathbf{a} + \overline{\mathbf{u}}_1 t, \overline{\mathbf{u}}_1)\,dt + \int_{0}^{s} \widehat{F}(\mathbf{a} + \overline{\mathbf{u}}_2 t, \overline{\mathbf{u}}_2)\,dt.$$

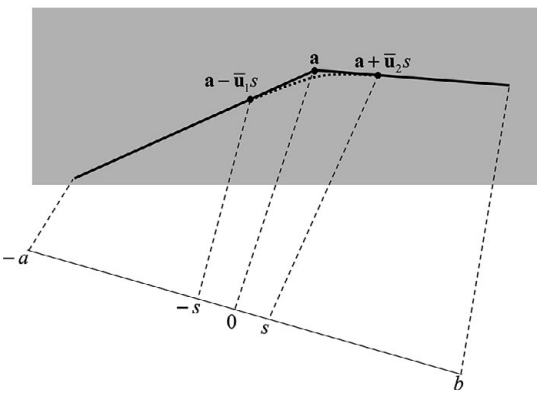


Figure 3.27 *An illustration for the corner-rounding procedure. The piecewise linear path is shown as a mapping of the interval* $[-a,b]$ *into Euclidean plane (gray area). At* 0 *the two segments meet, and around this point they are replaced by the path shown by the dotted line of an arbitrarily close length.*

Corner-rounding consists of replacing $\mathbf{p}|[-s,s]$ with a continuously differentiable path

$$\mathbf{q}(t) = \mathbf{a} + \mathbf{u}(t)\, t, \quad t \in [-s,s], \tag{3.41}$$

such that

$$\begin{array}{l} \mathbf{u}(-s) = \overline{\mathbf{u}}_1, \mathbf{u}(s) = \overline{\mathbf{u}}_2 \\ \dot{\mathbf{u}}(-s) = \dot{\mathbf{u}}(s) = \mathbf{0} \end{array} \tag{3.42}$$

and

$$\lim_{s \to 0+} D\mathbf{x}|[-s,s] = 0. \tag{3.43}$$

The requirements (3.42) ensure that the modified path $\mathbf{r}|[-a,b]$ defined by

$$\mathbf{r}(t) = \begin{cases} \mathbf{p}(t) & \text{if } t \notin [-s,s] \\ \mathbf{q}(t) & \text{if } t \in [-s,s] \end{cases}$$

is continuously differentiable. The requirement (3.43) ensures that the difference

$$|D\mathbf{p}|[-a,b] - D\mathbf{r}|[-a,b]|$$

can be made arbitrarily small by choosing s sufficiently small. One example of (3.41) is given by

$$\mathbf{u}(t) = \frac{\overline{\mathbf{u}}_1 + \overline{\mathbf{u}}_2}{2} + \frac{\left(\frac{t}{s}\right)^3 - 3\left(\frac{t}{s}\right)}{4}\left(\overline{\mathbf{u}}_1 - \overline{\mathbf{u}}_2\right).$$

We can now reformulate Theorem 3.56 as follows.

Theorem 3.57. *For every path* $\mathbf{h}|[a,b]$ *connecting* $\mathbf{a}$ *to* $\mathbf{b}$ *one can find a continuously differentiable path from* $\mathbf{a}$ *to* $\mathbf{b}$ *which is arbitrarily close to* $\mathbf{h}|[a,b]$ *pointwise and in its length.*

As an immediate consequence, we have the following.

Theorem 3.58. *If* G *in* $(\mathfrak{S}, D)$ *is an intrinsic metric, then, for any* $\mathbf{a}, \mathbf{b}$ *in* $\mathfrak{S}$*,*

$$G\mathbf{ab} = \inf \int_a^b \widehat{F}\left(\mathbf{f}(t), \dot{\mathbf{f}}(t)\right) dt,$$

where the infimum is taken across all continuously differentiable paths (or piecewise continuously differentiable, if more convenient) connecting $\mathbf{a}$ *to* $\mathbf{b}$*.*

Recall that G is defined as intrinsic if $G\mathbf{ab}$ is the infimum of the length of all paths connecting $\mathbf{a}$ to $\mathbf{b}$. This property is not derivable from the assumptions $\mathcal{E}1$ and $\mathcal{E}2$ we made about the relationship between $(\mathfrak{S}, D)$ and $(\mathfrak{S}, E)$. It should therefore be stipulated as an additional assumption or derived from other additional assumptions, such as that $(\mathfrak{S}, D)$ is a complete space with intermediate points.

3.9 Dissimilarity Cumulation: Extensions and Applications

In this section we give a few examples of extensions of the dissimilarity cumulation theory aimed at broadening the scope of its applicability.

3.9.1 Example 1: Observational Sorites "Paradox"

The issue of pairwise discrimination is the main application of Fechnerian scaling and the original motivation for its development. As we know from Sections 3.1.4 and 3.1.5, it is a fundamental fact that two stimuli being compared must belong to distinct observation areas, say, one being on the left and the other on the right in the visual field, or one being first and the other second in time. Without this one would not be able to speak, for example, of a stimulus with value $\mathbf{x}$ being compared to a stimulus with the same value, because then we would simply have a single stimulus. Similarly, without the distinct observation areas there would be no operational meaning in distinguishing $(\mathbf{x},\mathbf{y})$ from $(\mathbf{y},\mathbf{x})$. Throughout this chapter the observation areas in our notation were implicit: for example, we assumed that the stimulus written first in $(\mathbf{x},\mathbf{y})$ belongs to the first observation area, or that $\mathbf{x}$ always denotes a stimulus in the first observation area. Here, however, we will need to indicate observation areas explicitly: $\mathbf{v}^{(o)}$ means a stimulus with value $\mathbf{v}$ in observation area o. If we assume that the observation areas are fixed, we can denote them 1 and 2, so that every value $\mathbf{v}$ may be part of the stimuli $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$. Note that with this notation any pair $\left\{\mathbf{x}^{(1)},\mathbf{y}^{(2)}\right\}$ can be considered unordered, because $\left\{\mathbf{y}^{(2)},\mathbf{x}^{(1)}\right\}$ represents the same pair.

There is an apparent "paradox" related to pairwise comparisons that seems so compelling that many describe it as a well-known empirical fact. Quoting from R. Duncan Luce (1956):

> It is certainly well known from psychophysics that if "preference" is taken to mean which of two weights a person believes to be heavier after hefting them, and if "adjacent" weights are properly chosen, say a gram difference in a total weight of many grams, then a subject will be indifferent between any two "adjacent" weights. If indifference were transitive, then he would be unable to detect any weight differences, however great, which is patently false.

In other words, one can have a sequence of weights in which every two successive weights subjectively match each other, but the first and the last one do not. In philosophy, this seemingly paradoxical situation is referred to as *observational sorites*. The term "sorites" means "heap" in Greek, and the paradox is traced back to the Greek philosopher Eubulides (fourth century BCE). In fact, Eubulides dealt with another form of the paradox, one in which stimuli are mapped into one of two categories one at a time. This form of sorites requires a different analysis. In our case, we have pairs of stimuli mapped into categories "match" or "do not match." The resolution of this paradox is based on two considerations:

1. The relationship "$\mathbf{x}^{(1)}$ matches $\mathbf{y}^{(2)}$" (or vice versa) is computed from an ensemble of responses rather than observed as an individual response. Individual responses to the same pair $\left\{\mathbf{x}^{(1)}, \mathbf{y}^{(2)}\right\}$ vary, and the pair can only be associated with a probability of a response, say

$$\psi^*\left(\mathbf{x}^{(1)}, \mathbf{y}^{(2)}\right) = \Pr\left[\mathbf{x}^{(1)} \text{ is judged to be different from } \mathbf{y}^{(2)}\right]. \tag{3.44}$$

2. Stimuli $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ have the same value $\mathbf{v}$ but they are different. To repeat the same stimulus, it should be presented in the same observation area in addition to having the same value.

Applying these considerations to the above quotation from Luce, let

$$w_1, w_2, w_3, w_4, \ldots, w_n$$

be the sequence of weights about which Luce (and many others) think as one in which w_{k-1} and w_k match (for $k = 2, \ldots, n$) but w_1 and w_n do not. Such a sequence is called a *(comparative) soritical sequence*. Let us, however, assign these weights to observation areas, as they should be. One can, for instance, place one weight in an observer's left hand and another weight in her right hand to be hefted simultaneously, in which case $w^{(1)} = w^{(left)}$ and $w^{(2)} = w^{(right)}$. Or the observer can heft one weight first and the other weight after a short interval, in which case $w^{(1)} = w^{(first)}$ and $w^{(2)} = w^{(second)}$. Whichever the case, since two adjacent weights in our sequence are to be compared, they should belong to different observation areas:

$$w_1^{(1)}, w_2^{(2)}, w_3^{(1)}, w_4^{(2)}, \ldots, w_n^{(2)}.$$

The last and the first stimuli also should belong to different observation areas if they are to be compared, so n must be an even number. Assuming that the discrimination here is of the "greater–less" variety, we have a function

$$\gamma\left(x^{(1)}, y^{(2)}\right) = \Pr\left[x^{(1)} \text{ is judged to be lighter than } y^{(2)}\right],$$

and the match is determined by

$$\gamma\left(x^{(1)}, y^{(2)}\right) = \frac{1}{2}.$$

So we have

$$\gamma\left(w_1^{(1)}, w_2^{(2)}\right) = \gamma\left(w_2^{(2)}, w_3^{(1)}\right) = \gamma\left(w_3^{(1)}, w_4^{(2)}\right) = \cdots = \gamma\left(w_{n-1}^{(1)}, w_n^{(2)}\right) = \frac{1}{2}.$$

It is not obvious now that we can have $w_1 < w_2 < w_3 < w_4 < \cdots < w_n$. In fact, if we accept the usual model of a psychometric function γ, as in Figures 3.1 and 3.2, w_k is uniquely determined as a match for w_{k-1}, and, moreover:

$$\begin{aligned} w_1^{(1)} &= w_3^{(1)} = \cdots = w_{n-1}^{(1)}, \\ w_2^{(2)} &= w_4^{(2)} = \cdots = w_n^{(2)}. \end{aligned}$$

The sequence clearly is not soritical, because $w_1^{(1)}$ and $w_n^{(2)}$ (for an even n) necessarily match.

Generalizing, if one explicitly considers observation areas as part of stimuli's identity, the idea of soritical sequences becomes unfounded. If one further accepts the principles stipulated in Section 3.1.4, enabling one to construct a canonical space $(\mathfrak{S}, D)$, then soritical sequences become impossible. Essentially we are dealing with the problem of a reasonable definition of a match (PSE). We outline below an axiomatic scheme that defines stimulus spaces in which soritical sequences are impossible.

Not to be constrained to just two fixed observation areas, we consider a union of stimulus spaces indexed by observation areas:

$$\mathcal{S} = \bigcup_{\alpha \in \Omega} \mathfrak{S}_\omega^*.$$

We indicate the elements of $\mathfrak{S}_\omega^*$ by the corresponding superscript, say $\mathbf{x}^{(\omega)}$. The set $\mathcal{S}$ is endowed with a binary relation $\mathbf{x}^{(\alpha)}\mathrm{M}\mathbf{y}^{(\beta)}$ (read as "$\mathbf{x}$ in α is matched by $\mathbf{y}$ in β"). The most basic property of M is

$$\mathbf{x}^{(\alpha)}\mathrm{M}\mathbf{y}^{(\beta)} \Longrightarrow \alpha \neq \beta. \tag{3.45}$$

Definition 3.59. Given a space $(\mathcal{S}, \mathrm{M})$, we call a sequence $\mathbf{x}_1^{(\omega_1)}, \ldots, \mathbf{x}_n^{(\omega_n)}$ *well-matched* if

$$\omega_i \neq \omega_j \Longrightarrow \mathbf{x}_i^{(\omega_i)}\mathrm{M}\mathbf{x}_j^{(\omega_j)} \tag{3.46}$$

for all $i,j \in \{1, \ldots, n\}$. The stimulus space $(\mathcal{S}, \mathrm{M})$ is *well-matched* if, for any sequence $\alpha, \beta, \gamma \in \Omega$ and any $\mathbf{a}^{(\alpha)} \in \mathcal{S}$, there is a well-matched sequence $\mathbf{a}^{(\alpha)}, \mathbf{b}^{(\beta)}, \mathbf{c}^{(\gamma)}$.

In particular, in a well-matched space, for any $\mathbf{a}^{(\alpha)}$ and any $\beta \in \Omega$, one can find a $\mathbf{b}^{(\beta)} \in \mathcal{S}$ such that $\mathbf{a}^{(\alpha)}\mathrm{M}\mathbf{b}^{(\beta)}$ and $\mathbf{b}^{(\beta)}\mathrm{M}\mathbf{a}^{(\alpha)}$.

Definition 3.60. Two stimuli $\mathbf{a}^{(\omega)}, \mathbf{b}^{(\omega)}$ in $(\mathcal{S}, \mathrm{M})$ are called *equivalent*, in symbols $\mathbf{a}^{(\omega)}\mathrm{E}\mathbf{b}^{(\omega)}$, if for any $\mathbf{c}^{(\iota)} \in \mathcal{S}$,

$$\mathbf{c}^{(\iota)}\mathrm{M}\mathbf{a}^{(\omega)} \iff \mathbf{c}^{(\iota)}\mathrm{M}\mathbf{b}^{(\omega)}. \tag{3.47}$$

$(\mathcal{S}, \mathrm{M})$ is a *regular space* if, for any $\mathbf{a}^{(\omega)}, \mathbf{b}^{(\omega)}, \mathbf{c}^{(\omega')} \in \mathcal{S}$ with $\omega \neq \omega'$,

$$\mathbf{a}^{(\omega)}\mathrm{M}\mathbf{c}^{(\omega')} \wedge \mathbf{b}^{(\omega)}\mathrm{M}\mathbf{c}^{(\omega')} \Longrightarrow \mathbf{a}^{(\omega)}\mathrm{E}\mathbf{b}^{(\omega)}. \tag{3.48}$$

This is a generalization of the notion of psychological equality introduced in Section 3.1.4.

Definition 3.61. Given a space $(\mathcal{S}, \mathrm{M})$, a sequence $\mathbf{x}_1^{(\omega_1)}, \ldots, \mathbf{x}_n^{(\omega_n)}$ with $\mathbf{x}_i^{(\omega_i)} \in \mathcal{S}$ for $i = 1, \ldots, n$, is called *soritical* if

1. $\mathbf{x}_i^{(\omega_i)}\mathrm{M}\mathbf{x}_{i+1}^{(\omega_{i+1})}$ for $i = 1, \ldots, n-1$,
2. $\omega_1 \neq \omega_n$,
3. but it is not true that $\mathbf{x}_1^{(\omega_1)}\mathrm{M}\mathbf{x}_n^{(\omega_n)}$.

Well-matchedness and regularity can be shown to be independent properties. Our interest is in the spaces that are both regular and well-matched. It can be proved that

Theorem 3.62. *In a regular well-matched space it is impossible to form a soritical sequence.*

3.9.2 Example 2: Thurstonian-Type Representations

Consider now the special case of the regular well-matched spaces, when the matching (PSE) relation is defined through minima of a same–different discrimination probability function $\psi^\star : \mathfrak{S}_1^\star \times \mathfrak{S}_2^\star \to [0,1]$ in (3.44). The question we pose is whether $\psi^\star$ can be "explained" by a *random-utility* (or *Thurstonian*) *model*, according to which each stimulus is mapped into a random variable in some perceptual space, and the decision "same" or "different" is determined by the values of these random variables for the stimuli $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(2)}$.

Let us assume that both $\mathfrak{S}_1^\star \times \mathfrak{S}_2^\star$ are open connected regions of $\mathbb{R}^n$, and let us present the property of regular minimality (3.14) in the following special form: there is a homeomorphism $\mathbf{h} : \mathfrak{S}_1^\star \to \mathfrak{S}_2^\star$ (a continuous function with a continuous $\mathbf{h}^{-1}$) such that

$$\begin{cases} \arg\min_{\mathbf{y}} \psi^\star(\mathbf{x},\mathbf{y}) = \mathbf{h}(\mathbf{x}), \\ \arg\min_{\mathbf{x}} \psi^\star(\mathbf{x},\mathbf{y}) = \mathbf{h}^{-1}(\mathbf{y}). \end{cases} \tag{3.49}$$

Here we once again drop the superscripts in $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(2)}$. The function $\arg\min_{a_i} f(a_1,\ldots,a_n)$ indicates the value of the argument a_i at which f reaches its minimum (at fixed values of the remaining arguments). Empirical studies show that generally the *minimum-level function* $\psi^\star(\mathbf{x},\mathbf{h}(\mathbf{x}))$ *varies with* $\mathbf{x}$*:*

$$\psi^\star(\mathbf{x},\mathbf{h}(\mathbf{x})) \neq const. \tag{3.50}$$

Equivalently written,

$$\psi^\star\left(\mathbf{h}^{-1}(\mathbf{y}),\mathbf{y}\right) \neq const.$$

We call this property *nonconstant self-dissimilarity* of $\psi^\star$.

Rather than using regular minimality (3.49) to bring the stimulus space to a canonical form, we will use the following construction. Consider a point $(\mathbf{p},\mathbf{h}(\mathbf{p}))$ in $\mathfrak{S}_1^\star \times \mathfrak{S}_2^\star$ and a direction $\mathbf{u}$ in

$$\mathbb{U}^n = \left\{\mathbf{u} = \mathbf{x} - \mathbf{p} : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{p}\right\}.$$

For $(x,y) \in [-a,a]^2$, where a is a small positive number, the function

$$\lambda(x,y) = \psi^\star(\mathbf{p} + \mathbf{u}x, \mathbf{h}(\mathbf{p} + \mathbf{u}y))$$

is called a *patch* of the function $\psi^\star(\mathbf{x},\mathbf{y})$ at $(\mathbf{p},\mathbf{h}(\mathbf{p}))$. Note that the $(\mathbf{p},\mathbf{h}(\mathbf{p}))$ itself corresponds to $(x=0,y=0)$, and the graph of the PSE function $(\mathbf{x},\mathbf{h}(\mathbf{x}))$ in the vicinity of $\mathbf{x}=\mathbf{p}$ is mapped into the diagonal $\{(x,y) : x = y\}$. We have therefore

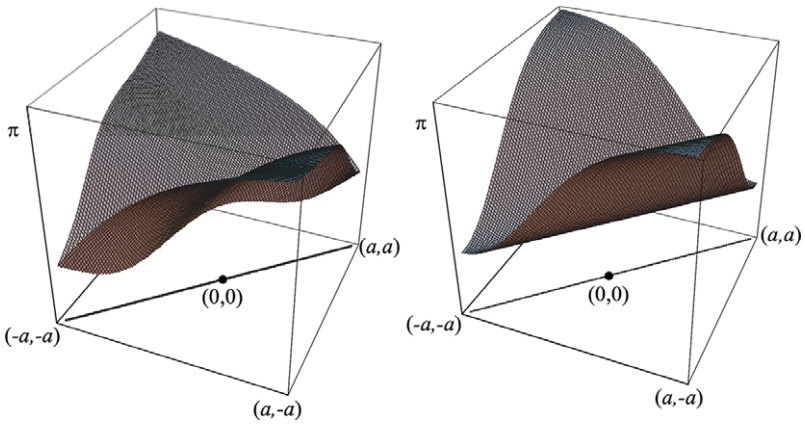


Figure 3.28 *A typical patch (left) and an atypical patch (right) on a small square* $[-a,a]^2$.

the following "patch-wise" version of the regular minimality and nonconstant self-dissimilarity:

$$\begin{cases} \arg\min_y \lambda\,(x,y) = x, \\ \arg\min_x \lambda\,(x,y) = y, \end{cases}$$

and

$$\lambda\,(x,x) \neq const.$$

for $(x,y) \in [-a,a]^2$. We will call a patch *typical* if $\lambda\,(x,x)$ is nonconstant for all sufficiently small positive a. Figure 3.28 illustrates the notion.

In a Thurstonian-type model (called so in honor of Leon Thurstone who introduced such models in psychology in the 1920s), there is some internal space of images P, and each stimulus $\mathbf{x} \in \mathfrak{S}_1^\star$ (hence also any x representing $\mathbf{x}$ in a patch) is mapped into a random variable A with values in P, and, similarly, $\mathbf{y} \in \mathfrak{S}_2^\star$ (hence also any y representing $\mathbf{y}$ in a patch) is mapped into a random variable B with values in P. We will denote these random variables $A\,(\mathbf{x})$ and $B\,(\mathbf{y})$, and their sets of possible values a and b, respectively. We will consider first the case when $A\,(\mathbf{x})$ and $B\,(\mathbf{y})$ are stochastically independent. According to the model, there is a function

$$d : \mathsf{a} \times \mathsf{b} \to \{\text{same, different}\},$$

determining which response will be given in a given presentation of the stimuli. In complete generality, with no constraints imposed, such a model is not falsifiable.

Theorem 3.63. *Any psychometric function* $\psi^\star : \mathfrak{S}_1^\star \times \mathfrak{S}_2^\star \to [0,1]$ *can be generated by a Thurstonian-type model with stochastically independent random variables* $A\,(\mathbf{x})$ *and* $B\,(\mathbf{y})$.

This is not, however, very interesting, because one normally would want to deal only with sufficiently "well-behaved" Thurstonian-type models. The intuition here is that, as $\mathbf{x}$ and $\mathbf{y}$ continuously change, the random variables $A\,(\mathbf{x})$ and $B\,(\mathbf{y})$ change sufficiently smoothly. Consider, for example, Figure 3.29, depicting a common way of modeling same–different comparisons. If the patch variables x and

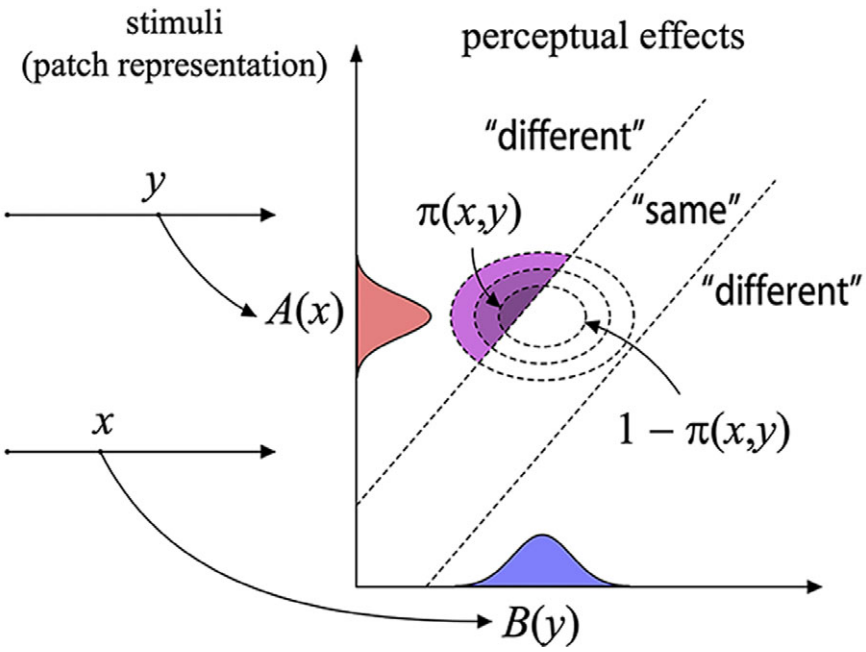


Figure 3.29 *A schematic representation of a Thurstonian-type model. The stimuli are represented by their patch variables x and y, and their perceptual effects are points in an interval of reals. The response "same" is given if and only if both random variables A (x) and B (y) fall within the area between the two dashed lines.*

y change by a small amount, one should expect that the shapes of the probability density functions do not change in an abrupt way. To formalize this intuition, denote, for any A-measurable set a in the perceptual space,

$$A_x(\mathsf{a}) = \Pr[A(x) \in \mathsf{a}],$$

and analogously, for any B-measurable set b in the perceptual space,

$$B_y(\mathsf{b}) = \Pr\left[B(y) \in \mathsf{b}\right].$$

Definition 3.64. Given a patch $\lambda(x, y)$, a Thurstonian-type model generating it is said to be *well-behaved* if, for every A-measurable set a and B-measurable set b, the left-hand and right-hand derivatives

$$\frac{dA_x(\mathsf{a})}{dx\pm}, \frac{dB_y(\mathsf{b})}{dy\pm}$$

exist, and are bounded across all measurable sets.

The latter means that there is a constant c such that

$$\left|\frac{dA_x(\mathsf{a})}{dx\pm}\right| < c, \left|\frac{dB_y(\mathsf{b})}{dy\pm}\right| < c$$

for all measurable a and b. The "textbook" distributions (such as normal, Weibull, etc.) with parameters depending on x and y in a piecewise differentiable way will always satisfy this definition.

Definition 3.65. A patch $\lambda\,(x,y)$ is called *near-smooth* if the left-hand and right-hand derivatives

$$\frac{\partial \lambda\,(x,y)}{\partial x\pm}$$

exist and are continuous in y; and similarly

$$\frac{\partial \lambda\,(x,y)}{\partial y\pm}$$

exist and are continuous in x.

It turns out that, perhaps not surprisingly,

Theorem 3.66. *A well-behaved Thurstonian representation can only generate near-smooth patches.*

A critical point in the development is created by the following fact.

Theorem 3.67. *No near-smooth patch can be typical, i.e., satisfy simultaneously the regular minimality and nonconstant self-dissimilarity properties.*

This means that for Thurstonian-type modeling of discrimination probabilities one cannot use well-behaved models, which in turn means the models should be quite complex mathematically (or else one should reject either regular minimality or nonconstant self-dissimilarity). With appropriate modifications of the definitions, this conclusion has been extended to Thurstonian models with *stochastically interdependent* (but *selectively influenced*) random variables, and to Thurstonian models in which the mapping of perceptual effects into responses is probabilistic too.

3.9.3 Example 3: Universality of Corrections for Violations of the Triangle Inequality

In Section 3.6 we described the Floyd–Warshall algorithm for finite stimulus spaces. It turns out that it can be extended to arbitrary sets, generally infinite and not necessarily discrete. This is done by using the Axiom of Choice of the set theory to index all triangles in a stimulus set by *ordinals*. An *ordinal* is a set α such that each $\beta \in \alpha$ is a set, and $\beta \subseteq \alpha$. Thus

$$\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \ldots \tag{3.51}$$

are (finite) ordinals. For any two ordinals α and β, one and only one of the following is true: $\alpha = \beta$, $\alpha \in \beta$, or $\beta \in \alpha$. The ordinals are ordered in the following way: if $\alpha \in \beta$, we write $\alpha < \beta$; if either $\alpha \in \beta$ or $\alpha = \beta$, we write $\alpha \leq \beta$. For each ordinal α, $\alpha \cup \{\alpha\}$ is also an ordinal, called the *successor* of α and denoted $\alpha + 1$. There are two types of ordinals:

1. *successor* ordinals α, such that α is the successor of another ordinal;
2. *limit* ordinals, those that do not succeed other ordinals.

Thus, we can identify $\emptyset$ in (3.51) with 0, and identify $n \cup \{n\}$ with $n+1$ for any ordinal identified with n. We have then that 0 is a limit ordinal, and each of $1, 2, 3, \ldots$ is a successor ordinal. The ordinal

$$\omega = \{0, 1, 2, 3, \ldots\}$$

is the smallest limit ordinal after 0, and the smallest infinite ordinal. The ordinals $\omega + 1, \omega + 2,$ etc. are again successor ordinals, $\omega + \omega$ is a limit ordinal, and so on. Theorems involving ordinals are often proved by *transfinite induction*: if a certain property holds for 0, and it holds for any ordinal α whenever it holds for all ordinals $\beta < \alpha$, then this property holds for all ordinals. Similarly, definitions of a property of ordinals can be given by means of *transfinite recursion*: if it is defined for 0, and if, having defined it for all $\beta < \alpha$, we can use our definition to define it for α, then we define it for all ordinals. Thus, in Definition 3.11, the procedure of correcting dissimilarity functions for violations of the triangle inequalities is described by means of the usual mathematical induction. It can be replaced with transfinite recursion as follows. We index the triangles **xyz** with pairwise distinct elements by ordinals, so that for every ordinal α there is an ordinal $\beta > \alpha$ indexing the same triangle. In other words, each triangle occurs an infinite number of times.

Definition 3.68. Define for each ordinal α a function $M^{(\alpha)} : \mathfrak{S} \times \mathfrak{S} \to \mathbb{R}$ as follows:

(i) $M^{(0)} \equiv D$;
(ii) for any successor ordinal $\alpha = \beta + 1$, and for all $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$,

$$M^{(\alpha)}\mathbf{ab} = \begin{cases} \min\{M^{(\beta)}\mathbf{ab}, M^{(\beta)}\mathbf{ax} + M^{(\beta)}\mathbf{xb}\} & \text{if } \mathbf{axb} \text{ is indexed by } \beta, \\ M^{(\beta)}\mathbf{ab} & \text{otherwise;} \end{cases}$$

(iii) if α is a limit ordinal, then, for all $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$,

$$M^{(\alpha)}\mathbf{ab} = \inf_{\beta < \alpha} M^{(\beta)}\mathbf{ab}.$$

It turns out that all results presented in Section 3.6 have their transfinite analogous in this generalization. In particular, "eventually" (i.e., at some ordinal α) the procedure is terminated with $M^{(\alpha)}$ coinciding with the quasimetric dissimilarity G, as defined in (3.18).

3.9.4 Example 4: Data Analysis

Multidimensional scaling (MDS) and clustering are among the widely used tools of data analysis and data visualization. The departure point of MDS is a matrix

$$\left\{d_{ij} : i, j = 1, 2, \ldots, n\right\}$$

whose entries are values of a dissimilarity function on the set of objects $\mathfrak{S} = \{1, 2, \ldots, n\}$. This requires that, for all $i \neq j$,

$$d_{ii} = 0 \text{ and } d_{ij} > 0.$$

If this is not the case, but regular minimality is satisfied, the matrix can be brought first to a canonical form, so that d_{ii} is the smallest value both in the ith row and in the ith column. Then one can replace d_{ij} with

$$\delta_{ij}^{(1)} = d_{ij} - d_{ii},$$

or with

$$\delta_{ij}^{(2)} = d_{ji} - d_{ii}.$$

The choice between the two corresponds to the choice between psychometric increments of the first and second kind. We know that this choice is immaterial in Fechnerian scaling, but in MDS it is immaterial only if the matrix is symmetrical:

$$d_{ij} = d_{ji}.$$

If this is not the case, one usually uses in MDS some symmetrization procedure: for example, one can replace each d_{ij} with

$$\delta_{ij} = d_{ij} + d_{ji} - d_{ii} - d_{jj} = \begin{cases} \delta_{ij}^{(1)} + \delta_{ji}^{(1)} \\ \delta_{ij}^{(2)} + \delta_{ji}^{(2)} \end{cases},$$

proposed by Roger Shepard in the 1950s for so-called *confusion matrices* (we will refer to it as *Shepard symmetrization*, SS). Following these or similar modifications, the matrix δ_{ij} can be viewed as a symmetric dissimilarity function.

If in addition the entries of the matrix satisfy the triangle inequality, the matrix represents a true metric on the set $\mathfrak{S} = \{1, 2, \ldots, n\}$. In such a case one can apply a procedure of *metric* MDS (mMDS), that consists in embedding the n elements of $\mathfrak{S}$ in an $\mathbb{R}^k$ so that the distances Δ_{ij} between the points are as close as possible to the corresponding δ_{ij}. The quality of approximation is usually estimated by a measure called *stress*, one variant of which is

$$\left(\frac{\sum_{i,j} \left(\Delta_{ij} - \delta_{ij} \right)^2}{\sum_{i,j} \delta_{ij}^2} \right)^{1/2}.$$

Since one of the goals of MDS is to help one to visualize the data, the distance in $\mathbb{R}^k$ is usually chosen to be Euclidean, and k chosen as small as possible (preferably 2 or 3).

However, in most applications δ_{ij} does not satisfy the triangle inequality, because of which MDS is used in its *nonmetric* version (nmMDS): here one seeks an embedding into a low-dimensional $\mathbb{R}^k$ in which the Euclidean distances match as closely as possible not δ_{ij} but some monotonically increasing transformation of δ_{ij}. The stress measure then has the form

$$\left(\frac{\sum_{i,j} \left(\Delta_{ij} - g\left(\delta_{ij} \right) \right)^2}{\sum_{i,j} g\left(\delta_{ij}^2 \right)} \right)^{1/2},$$

minimized across all possible monotone functions g.

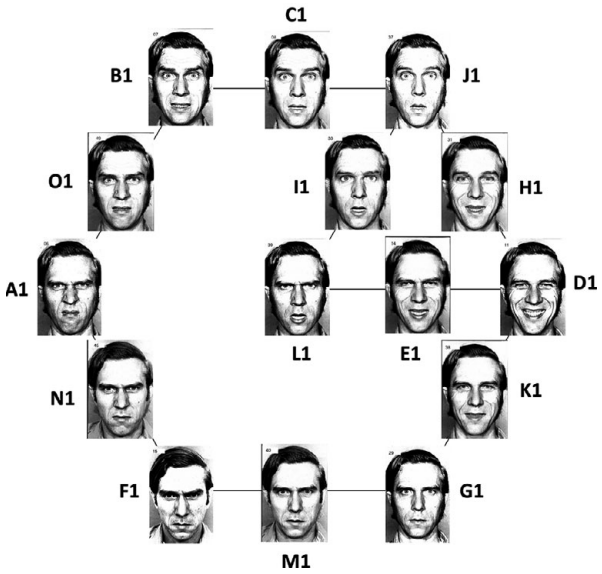


Figure 3.30 *A sample of faces presented two at a time with the question whether they represent the same emotion or different emotions.*

Dissimilarity cumulation offers a different approach to the same problem, one that does not require any transformations. Once the original matrix d_{ij} is brought to a canonical form and replaced with $\delta_{ij}^{(1)}$ or $\delta_{ij}^{(2)}$, one computes from either of them the Fechnerian distances $\overleftrightarrow{G}_{ij}$. Since these are true distances, one can apply to them the metric version of MDS to seek a low-dimensional Euclidean embedding. For illustration, consider an experiment reported in Dzhafarov and Paramei (2010). Images of faces shown in Figure 3.30 were presented two at a time, and the observer was asked to determine whether they exhibited the same emotion or different emotions. The data d_{ij} were estimates of the probabilities of the response "different emotions." Figure 3.31 shows the value of stress as a function of k in the embedding space $\mathbb{R}^k$ (so-called *scree plots*). The comparison of the two procedures

(DC-mMDS) metric MDS applied to the results of dissimilarity cumulation, and
(SS-nmMDS) nonmetric MDS applied to Shepard-symmetrized data

shows that the former seems to better identify the minimal dimensionality of the embedding space. In DC-mMDS, an acceptably small value of stress is achieved at $k = 2$ or 3, and stress drops very slowly afterwards, whereas in SS-nmMDS, the deceleration of the scree plot is less pronounced. Having chosen, say, $k = 3$, the results of both procedures can be further subjected to cluster analysis, which groups the points in $\mathbb{R}^3$ into a designated number of *clusters* (the K-means procedure) or constructs their *dendrogram* (hierarchical cluster analysis). We do not discuss these procedure, as our goal is to merely point out that Fechnerian scaling allows one to base all of them on true distances, without resorting to

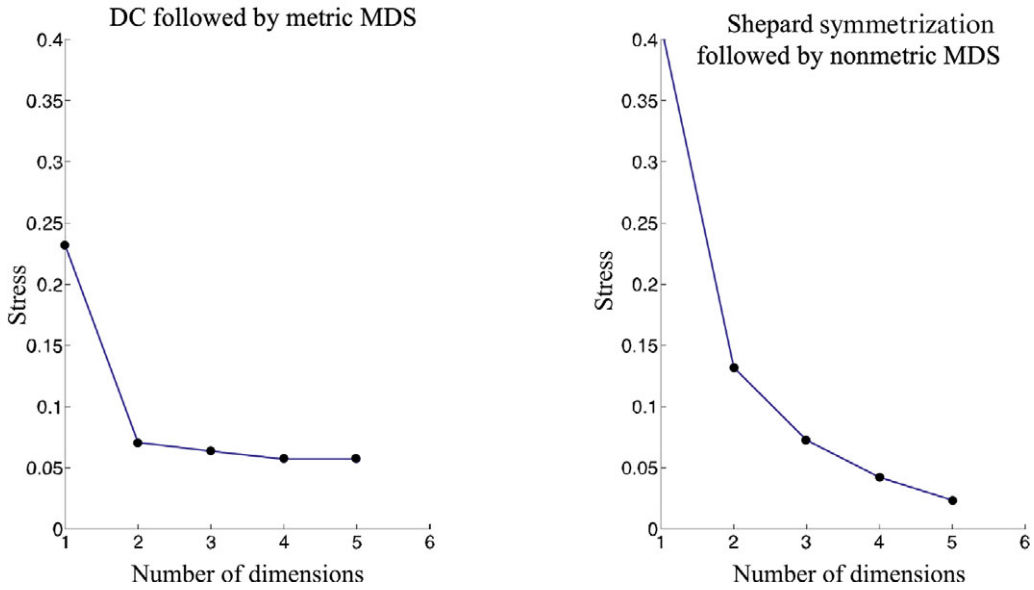


Figure 3.31 *Scree plots of mMDS following Fechnerian scaling (left) and nmMDS following Shepard's symmetrization. The optimal number of dimensions is usually chosen as one at which the scree plot visibly decelerates (exhibits a "knee").*

an unconstrained search of a monotone transformation. The example in the next section describes an alternative to the dissimilarity cumulation approach that results in a cluster analysis representation.

There are two public-domain programs that perform MDS and clustering of the results of dissimilarity cumulation. One of them is the Matlab-based software package FSCAMDS (stands for *F*echnerian *S*caling – *C*lustering – *a*nd – *M*ulti*d*imensional *S*caling), the other is the R-language package fechner (see the next section for references). These data-analytic programs have a variety of options, of which we will mention the following.

It is sometimes the case, especially if the data are probabilities, or if they are sampled from a path-connected space, that large values of dissimilarity are unreliable, and the cumulation is to be restricted only to smaller values. The software packages allow one to set a value above which a dissimilarity $D\mathbf{ab}$ is replaced with infinity, removing thereby the link **ab** from the cumulation process (because the latter seeks the smallest cumulated value).

It is sometimes the case that regular minimality in the original data set is violated. The software packages allow one to choose between the following options:

1. to "doctor" the data by designating the pairs of PSE and, following the canonical transformation, to replace negative values of $d_{ij} - d_{ii}$ with zero;

2. to perform Fechnerian scaling separately for the two observation areas, obtaining thereby $\overleftrightarrow{G_1}$ and $\overleftrightarrow{G_2}$ distances, not equal to each other.

The justifiability of the second option depends on one's position with respect to the empirical status of the regular minimality law. As mentioned in Section 3.1.5, regular minimality in this chapter is not taken as an empirical claim. Rather, it has been part of the definition of the functions we have dealt with in our mathematical theory.

3.9.5 Example 5: Ultrametric Fechnerian Scaling

There is a more direct way to obtain a representation of dissimilarities by hierarchical clusters (dendrogram or *rooted tree*). The basic idea consists in replacing "dissimilarity cumulation" by a "dissimilarity maximization" procedure.

Given a chain $\mathbf{X} = \mathbf{x}_1 \ldots \mathbf{x}_n$ and a binary (real-valued) function F, the notation $\Delta_F \mathbf{X}$ stands for

$$\max_{i=1,\ldots,n-1} F\mathbf{x_i x_{i+1}},$$

again with the obvious convention that the quantity is zero if n is 1 or 0. A dissimilarity function M on a finite set $\mathfrak{S}$ is called a *quasi-ultrametric* if it satisfies the *ultrametric inequality*

$$\max\{M\mathbf{ab}, M\mathbf{bc}\} \geq M\mathbf{ac} \tag{3.52}$$

for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathfrak{S}$.

The ultrametric inequality is rather restrictive: it is equivalent to postulating that, for any triple of elements, two dissimilarities have to be equal and not smaller than the third.

Definition 3.69. Given a dissimilarity D on a finite set $\mathfrak{S}$, the *quasi-ultrametric* G^∞ *induced by* D is defined as

$$G^\infty \mathbf{ab} = \min_{\mathbf{X} \in \mathcal{C}} \Delta_D \mathbf{aXb}, \tag{3.53}$$

for all $\mathbf{a}, \mathbf{b} \in \mathfrak{S}$.

Thus, the value of $G^\infty \mathbf{ab}$ is obtained by taking the minimum, across all chains $\mathbf{X}$ from $\mathbf{a}$ to $\mathbf{b}$, of the maximum dissimilarity value of the chain. That G^∞ is a quasi-ultrametric is easy to prove. A reasonable symmetrization procedure, yielding a metric, is

$$G^{\infty^*} \mathbf{ab} = \max\{G^\infty \mathbf{ab}, G^\infty \mathbf{ba}\}, \tag{3.54}$$

called the *overall Fechnerian ultrametric* on $\mathfrak{S}$.

The ultrametric inequality is often violated in empirical data. However, in analogy to recursive corrections for violations of the triangle inequality, it can be shown that a corresponding series of recursive corrections on the dissimilarity

values for violations of the ultrametric inequality would yield the induced quasi-ultrametric distances. This is in contrast to applying the different standard hierarchical cluster algorithms (like single-link, combined-link, etc.) to one and the same data set: when violations exist, these algorithms will typically result in rather different ultrametrics.

One can consider procedures intermediate between cumulation and maximization of dissimilarities by defining, for any dissimilarity function D, the length of a chain $\mathbf{X} = \mathbf{x}_1, \ldots \mathbf{x}_n$ by

$$D\mathbf{X} = ((D\mathbf{x_1x_2})^k + \cdots + (D\mathbf{x}_{n-1}\mathbf{x}_n)^k)^{1/k}. \tag{3.55}$$

For $k \to \infty$ this would result in the ultrametric approach outlined above. For finite k, the procedure is generalizable to arbitrary dissimilarity spaces. This follows from the fact that the use of (3.55) is equivalent to the use of the original dissimilarity cumulation procedure in which one first redefines D into D^k (which yields another dissimilarity function), and then redefines the quasimetric G induced by D^k into $G^{1/k}$ (which yields another quasimetric).

3.10 Related Literature

Fechner's original theory is presented in the *Elemente der Psychophysik* (Fechner, 1860), but important additions and clarifications can be found in a later book (Fechner, 1877), and in a paper written shortly before Fechner's death (Fechner, 1887). A detailed modern account of Fechner's original theory, especially the ways he derived his logarithmic psychophysical law, can be found in Dzhafarov and Colonius (2011). For related interpretations of Fechner's theory, see Creelman (1967), Falmagne (1971), Krantz (1971), and Pfanzagl (1962). A different interpretation of Fechner's theory, one that finds it lacking in mathematical coherence and with which we disagree, is presented in Luce and Edwards (1958) and Luce and Galanter (1963).

The theory of dissimilarity cumulation is presented in Dzhafarov and Colonius (2007) and elaborated in Dzhafarov (2008a) (see also Dzhafarov 2009). The geometric aspects of this theory are close to those of the distance and geodesics theory developed in Blumenthal (1953), Blumenthal and Menger (1970), and Busemann (2005). To better understand the topology and uniformity aspects of dissimilarity cumulation, one can consult, for example, Hocking and Young (1961) and Kelly (1955). A proof of Theorem 3.10 can be found in Dzhafarov and Colonius (2007). A proof of Theorem 3.26 is presented in Dzhafarov (2008a).

For stimuli spaces defined on regions of $\mathbb{R}^n$, the mathematical theory essentially becomes a generalized form of Finsler geometry, as presented in Dzhafarov (2008b). A more detailed presentation, however, and one closer to this chapter, is found in earlier work (Dzhafarov & Colonius, 1999, 2001). This part of the theory has its precursors in Helmholtz (1891) and Schrödinger (1920/1970, 1926/1970), both of whom, in different ways, used Fechner's cumulation of

infinitesimal differences to construct a Riemannian geometry (a special case of Finsler geometry) of color space.

In this chapter we have entirely omitted the important topic of invariance of length and distance under homeomorphic (for general path-connected spaces) and diffeomorphic (for $\mathbb{R}^n$-based spaces) transformations of space and reparameterizations of paths. These topics are discussed in Dzhafarov (2008b) and Dzhafarov and Colonius (2001). We have also ignored the difference between paths and arcs, discussed in detail in Dzhafarov (2008b).

Dissimilarity cumulation in discrete stimulus spaces is described in Dzhafarov (2010a) and Dzhafarov and Colonius (2006a, 2006b). The generalization of the Floyd–Warshall algorithm to arbitrary spaces (Section 3.9.3) is described in D. D. Dzhafarov and Dzhafarov (2011).

The notion of separate observation area in stimulus comparisons, as well as the regular minimality law, have been initially formulated in Dzhafarov (2002) and elaborated in Kujala and Dzhafarov (2008, 2009a). The application of the regularity and well-matchedness principles to the comparative sorites "paradox" is presented in Dzhafarov and Dzhafarov (2010), with a proof of Theorem 3.62, and in Dzhafarov and Perry (2014).

The application of these principles, together with nonconstant self-dissimilarity to Thurstonian-type modeling, is presented in Dzhafarov (2003a, 2003b), where one can find proofs of the theorems in Section 3.9.2. This part of the theory has been generalized and greatly extended in Kujala and Dzhafarov (2008, 2009a, 2009b).

For multidimensional scaling see, for example, Borg and Groenen (1997). Clustering procedures, hierarchical and K-means, are described in standard textbooks of multivariate statistics (e.g., Everitt *et al.*, 2011). The ultrametric Fechnerian scaling approach is presented in Colonius and Dzhafarov (2012).

The link and instructions to the R language software package fechner mentioned in Section 3.9.4 are available in Ünlü, Kiefer, and Dzhafarov (2009). The link and instructions to the software package FSCAMDS are available in Dzhafarov (2010).

Appendix: Select Proofs

Theorem 3.30. *$F(\mathbf{x},\mathbf{u})$ is well-defined for any $(\mathbf{x},\mathbf{u}) \in \mathbb{T} \cup \{(\mathbf{x},\mathbf{0}) : \mathbf{x} \in \mathfrak{S}\}$. It is positive for $\mathbf{u} \neq \mathbf{0}$, continuous in $(\mathbf{x},\mathbf{u})$, and Euler homogeneous in $\mathbf{u}$.*

Proof. We first show that $F(\mathbf{x},\overline{\mathbf{u}})$ is continuous in $(\mathbf{x},\overline{\mathbf{u}})$. By Assumptions $\mathcal{E}2$, for any $\varepsilon > 0$ there is a $\delta = \delta(\mathbf{x},\overline{\mathbf{u}},\varepsilon) > 0$ such that

$$\max\left\{|\mathbf{a}-\mathbf{x}|, |\mathbf{b}-\mathbf{x}|, \left|\overline{\mathbf{b}-\mathbf{a}} - \overline{\mathbf{u}}\right|\right\} < \delta(\mathbf{x},\overline{\mathbf{u}},\varepsilon) \Longrightarrow \left|\frac{D\mathbf{ab}}{|\mathbf{b}-\mathbf{a}|} - F(\mathbf{x},\overline{\mathbf{u}})\right| < \varepsilon.$$

Consider a sequence $(\mathbf{x}_n,\overline{\mathbf{u}}_n) \to (\mathbf{x},\overline{\mathbf{u}})$, and let $(\mathbf{a}_n,\mathbf{b}_n)$, $\mathbf{a}_n \neq \mathbf{b}_n$, be any sequence satisfying

$$\max\left\{|\mathbf{a}_n-\mathbf{x}_n|, |\mathbf{b}_n-\mathbf{x}_n|, \left|\overline{\mathbf{b}_n-\mathbf{a}_n} - \overline{\mathbf{u}}_n\right|\right\} < \min\left\{\delta\left(\mathbf{x}_n,\overline{\mathbf{u}}_n,\frac{1}{n}\right), \frac{1}{2}\delta(\mathbf{x},\overline{\mathbf{u}},\varepsilon)\right\}.$$

Clearly

$$\frac{D\mathbf{a}_n\mathbf{b}_n}{|\mathbf{b}_n-\mathbf{a}_n|}-F\left(\mathbf{x}_n,\overline{\mathbf{u}}_n\right)\to 0.$$

At the same time, for all sufficiently large n,

$$\max\left\{|\mathbf{x}_n-\mathbf{x}|\,,\,|\overline{\mathbf{u}}_n-\overline{\mathbf{u}}|\right\}<\frac{1}{2}\delta\left(\mathbf{x},\overline{\mathbf{u}},\varepsilon\right),$$

implying

$$\max\left\{|\mathbf{a}_n-\mathbf{x}|\,,\,|\mathbf{b}_n-\mathbf{x}|\,,\,\left|\overline{\mathbf{b}_n-\mathbf{a}_n}-\overline{\mathbf{u}}\right|\right\}<\delta\left(\mathbf{x},\overline{\mathbf{u}},\varepsilon\right).$$

But then

$$\left|\frac{D\mathbf{a}_n\mathbf{b}_n}{|\mathbf{b}_n-\mathbf{a}_n|}-F\left(\mathbf{x},\overline{\mathbf{u}}\right)\right|<\varepsilon,$$

and, as ε can be chosen arbitrarily small, we have

$$\frac{D\mathbf{a}_n\mathbf{b}_n}{|\mathbf{b}_n-\mathbf{a}_n|}-F\left(\mathbf{x},\overline{\mathbf{u}}\right)\to 0.$$

The convergence

$$F\left(\mathbf{x}_n,\overline{\mathbf{u}}_n\right)\to F\left(\mathbf{x},\overline{\mathbf{u}}\right)$$

follows, establishing the continuity of $F\left(\mathbf{x},\overline{\mathbf{u}}\right)$. Now, for $\mathbf{u}\neq\mathbf{0}$, denoting $\mathbf{u}=|\mathbf{u}|\,\overline{\mathbf{u}}$,

$$F\left(\mathbf{x},\mathbf{u}\right)=\lim_{s\to 0+}\frac{D\mathbf{x}\left[\mathbf{x}+\mathbf{u}s\right]}{s}=|\mathbf{u}|\lim_{|\mathbf{u}|s\to 0+}\frac{D\mathbf{x}\left[\mathbf{x}+\overline{\mathbf{u}}\,|\mathbf{u}|\,s\right]}{|\mathbf{u}|\,s}=|\mathbf{u}|\,F\left(\mathbf{x},\overline{\mathbf{u}}\right).$$

It immediately follows that $F\left(\mathbf{x},\mathbf{u}\right)$ exists, that it is positive and continuous, and that

$$F\left(\mathbf{x},\mathbf{u}\right)=|\mathbf{u}|\,F\left(\mathbf{x},\overline{\mathbf{u}}\right).$$

So, for $k>0$,

$$F\left(\mathbf{x},k\mathbf{u}\right)=k\,|\mathbf{u}|\,F\left(\mathbf{x},\overline{\mathbf{u}}\right)=kF\left(\mathbf{x},\mathbf{u}\right).$$

Finally, since any convergence of $(\mathbf{x}_n,\mathbf{u}_n)\to(\mathbf{x},\mathbf{0})$ with $\mathbf{u}_n\neq\mathbf{0}$ can be presented as $(\mathbf{x}_n,\,|\mathbf{u}_n|\,\overline{\mathbf{u}}_n)\to(\mathbf{x},\mathbf{0})$ with $|\mathbf{u}_n|\to 0$, we have

$$F\left(\mathbf{x}_n,\mathbf{u}_n\right)=|\mathbf{u}_n|\,F\left(\mathbf{x}_n,\overline{\mathbf{u}}_n\right)\to 0,$$

because within a small ball around $\mathbf{x}$ and on a compact set of unit vectors the function $F\left(\mathbf{x}_n,\overline{\mathbf{u}}_n\right)$ does not exceed some finite value. Thus $F\left(\mathbf{x}_n,\mathbf{u}_n\right)$ extends to $F\left(\mathbf{x},\mathbf{0}\right)=0$ by continuity. □

Lemma 3.42. *For any* $(\mathbf{a},\mathbf{u})\in\mathbb{T}$, *the maximal production of* $\mathbf{u}$ *in* $\mathbb{I}_{\mathbf{a}}$ *can be presented as a convex combination of n (not necessarily distinct) radius-vectors* $\mathbf{v}_1,\ldots,\mathbf{v}_n\in\delta\mathbb{I}_{\mathbf{a}}$.

Proof. With no loss of generality, let $\mathbf{u} \in \delta\mathbb{I}_{\mathbf{a}}$, and let κ stand for $\kappa(\mathbf{a}, \mathbf{u})$. By Corollary 3.36, for some $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1} \in \mathbb{I}_{\mathbf{a}}$, the system of $n+1$ linear equations

$$\begin{cases} \kappa\mathbf{u} = \lambda_1\mathbf{v}_1 + \cdots + \lambda_{n+1}\mathbf{v}_{n+1} \\ \lambda_1 + \cdots + \lambda_{n+1} = 1 \end{cases}$$

has a solution $\lambda_1, \ldots, \lambda_{n+1} \in [0,1]$. Assume that $\lambda_1, \ldots, \lambda_{n+1}$ are all positive (if some of them are zero, the theorem's statement holds). If the determinant of the matrix of coefficients for this system were nonzero, then, for any ε, the modified system

$$\begin{cases} [\kappa + \varepsilon]\,\mathbf{u} = \lambda_1\mathbf{v}_1 + \cdots + \lambda_{n+1}\mathbf{v}_{n+1} \\ \lambda_1 + \cdots + \lambda_{n+1} = 1 \end{cases}$$

would also have a solution $\lambda'_1, \ldots, \lambda'_{n+1}$, and choosing ε positive and sufficiently small, this solution (by continuity) would also satisfy $\lambda'_1 > 0, \ldots, \lambda'_{n+1} > 0$. But this would mean that $[\kappa + \varepsilon]\,\mathbf{u}$ belongs to the convex hull of $\mathbb{I}_{\mathbf{a}}$, which is impossible since $\kappa\mathbf{u}$ is the maximal production of $\mathbf{u}$. Hence

$$\det\begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_{n+1} \\ 1 & \cdots & 1 \end{bmatrix} = 0,$$

where we treat $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1}$ as n-element columns. But this means that, for some $\gamma_1, \ldots, \gamma_{n+1}$, not all zero,

$$\gamma_1\begin{bmatrix} \mathbf{v}_1 \\ 1 \end{bmatrix} + \cdots + \gamma_{n+1}\begin{bmatrix} \mathbf{v}_{n+1} \\ 1 \end{bmatrix} = \mathbf{0},$$

which indicates the affine dependence of $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1}$. It follows from Lemma 3.35 that $\mathbf{u}$ can be presented as a convex combination of some $m < n+1$ (not necessarily distinct) nonzero vectors in $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1} \in \mathbb{I}_{\mathbf{a}}$. Let them be the first m vectors in the list, $\mathbf{v}_1, \ldots, \mathbf{v}_m$. We now have the system

$$\begin{cases} \kappa\mathbf{u} = \lambda_1\mathbf{v}_1 + \cdots + \lambda_m\mathbf{v}_m \\ \lambda_1 + \cdots + \lambda_m = 1 \end{cases}$$

with a solution $\lambda_1 > 0, \ldots, \lambda_m > 0$ (zero values here would simply decrease m). Rewriting it as

$$\begin{cases} \kappa\mathbf{u} = \lambda_1 c_1\widetilde{\mathbf{v}}_1 + \cdots + \lambda_m c_m\widetilde{\mathbf{v}}_m \\ \lambda_1 + \cdots + \lambda_m = 1 \end{cases},$$

where $\widetilde{\mathbf{v}}_i \in \delta\mathbb{I}_{\mathbf{a}}$ is codirectional with $\mathbf{v}_i$ $(i = 1, \ldots, m)$, it is clear by Lemma 3.34 that for κ to have a maximal possible value, all c_i should have maximal possible values. In $\mathbb{I}_{\mathbf{a}}$ these values are $c_1 = \cdots = c_m = 1$, that is, all vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are radius-vectors. This completes the proof. □

Theorem 3.46. *The minimal submetric function* $\widehat{F}(\mathbf{a}, \mathbf{u})$ *has all the properties of a submetric function: it is positive for* $\mathbf{u} \neq \mathbf{0}$*, Euler homogeneous, and continuous.*

Proof. We only prove the continuity, as the other properties follow trivially from the definition of $\widehat{F}$ and the analogous properties of F. Consider a sequence of line elements

$$(\mathbf{a}_k, \mathbf{u}_k) \to (\mathbf{a}, \mathbf{u}) .$$

Let $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ be a minimizing chain for $(\mathbf{a}, \mathbf{u})$ (or a sequence of n zero vectors if $\mathbf{u} = \mathbf{0}$). For every k, consider the sequence $\mathbf{v}_1 + (\mathbf{u}_k - \mathbf{u}), \mathbf{v}_2, \ldots, \mathbf{v}_n$, which differs from the minimizing chain in the first element only. Its elements sum to $\mathbf{u}_k$, because of which

$$F(\mathbf{a}_k, \mathbf{v}_1 + (\mathbf{u}_k - \mathbf{u})) + F(\mathbf{a}_k, \mathbf{v}_2) + \cdots + F(\mathbf{a}_k, \mathbf{v}_n) \geq \widehat{F}(\mathbf{a}_k, \mathbf{u}_k) .$$

At the same time, by continuity of F,

$$\begin{aligned} &F(\mathbf{a}_k, \mathbf{v}_1 + (\mathbf{u}_k - \mathbf{u})) + F(\mathbf{a}_k, \mathbf{v}_2) + \cdots + F(\mathbf{a}_k, \mathbf{v}_n) \\ &\to F(\mathbf{a}_k, \mathbf{v}_1) + F(\mathbf{a}_k, \mathbf{v}_2) + \cdots + F(\mathbf{a}_k, \mathbf{v}_n) = \widehat{F}(\mathbf{a}, \mathbf{u}) , \end{aligned}$$

whence it follows that

$$\limsup_{k\to\infty} \widehat{F}(\mathbf{a}_k, \mathbf{u}_k) \leq \widehat{F}(\mathbf{a}, \mathbf{u}) .$$

To prove that at the same time

$$\liminf_{k\to\infty} \widehat{F}(\mathbf{a}_k, \mathbf{u}_k) \geq \widehat{F}(\mathbf{a}, \mathbf{u}) ,$$

let $(\mathbf{v}_{1k}, \ldots, \mathbf{v}_{nk})$ be a minimizing chain for $(\mathbf{a}_k, \mathbf{u}_k)$, for every k, and consider the sequence $\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k), \mathbf{v}_{2k}, \ldots, \mathbf{v}_{nk}$, which differs from the minimizing chain in the first element only. Its elements sum to $\mathbf{u}$, because of which

$$F(\mathbf{a}, \mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)) + F(\mathbf{a}, \mathbf{v}_{2k}) + \cdots + F(\mathbf{a}, \mathbf{v}_{nk}) \geq \widehat{F}(\mathbf{a}, \mathbf{u}) .$$

We will arrive at the desired inequality for lim inf if we show that

$$[F(\mathbf{a}, \mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)) + F(\mathbf{a}, \mathbf{v}_{2k}) + \cdots + F(\mathbf{a}, \mathbf{v}_{nk})] - \widehat{F}(\mathbf{a}_k, \mathbf{u}_k) \to 0.$$

The left-hand side difference here is

$$\begin{aligned} &[F(\mathbf{a}, \mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)) + F(\mathbf{a}, \mathbf{v}_{2k}) + \cdots + F(\mathbf{a}, \mathbf{v}_{nk})] \\ &\quad - [F(\mathbf{a}_k, \mathbf{v}_{1k}) + F(\mathbf{a}_k, \mathbf{v}_{2k}) + \cdots + F(\mathbf{a}_k, \mathbf{v}_{nk})] \\ &= [F(\mathbf{a}, \mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)) - F(\mathbf{a}_k, \mathbf{v}_{1k})] + [F(\mathbf{a}, \mathbf{v}_{2k}) - F(\mathbf{a}_k, \mathbf{v}_{2k})] \\ &\quad + \cdots + [F(\mathbf{a}, \mathbf{v}_{nk}) - F(\mathbf{a}_k, \mathbf{v}_{nk})] , \end{aligned}$$

where

$$\begin{aligned} &F(\mathbf{a}, \mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)) - F(\mathbf{a}_k, \mathbf{v}_{1k}) \\ &\quad = (|\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)| - |\mathbf{v}_{1k}|) F\left(\mathbf{a}, \overline{\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)}\right) \\ &\qquad + |\mathbf{v}_{1k}| \left[F\left(\mathbf{a}, \overline{\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)}\right) - F(\mathbf{a}_k, \overline{\mathbf{v}}_{1k})\right] , \end{aligned}$$

and

$$F(\mathbf{a}, \mathbf{v}_{ik}) - F(\mathbf{a}_k, \mathbf{v}_{ik}) = |\mathbf{v}_{ik}| [F(\mathbf{a}, \overline{\mathbf{v}}_{ik}) - F(\mathbf{a}_k, \overline{\mathbf{v}}_{ik})] , i = 2, \ldots, n.$$

Since $\mathbf{u}_k \to \mathbf{u}$, $\mathbf{a}_k \to \mathbf{a}$, and F is uniformly continuous and bounded on the compact set of unit vectors, we have

$$\begin{gathered}
|\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)| - |\mathbf{v}_{1k}| \to 0, \\
F\left(\mathbf{a}, \overline{\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)}\right) - F\left(\mathbf{a}_k, \overline{\mathbf{v}}_{1k}\right) \to 0, \\
\left(|\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)| - |\mathbf{v}_{1k}|\right) F\left(\mathbf{a}, \overline{\mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)}\right) \to 0, \\
F\left(\mathbf{a}, \overline{\mathbf{v}}_{ik}\right) - F\left(\mathbf{a}_k, \overline{\mathbf{v}}_{ik}\right) \to 0.
\end{gathered}$$

To see that

$$\begin{gathered}
F\left(\mathbf{a}, \mathbf{v}_{1k} + (\mathbf{u} - \mathbf{u}_k)\right) - F\left(\mathbf{a}_k, \mathbf{v}_{1k}\right) \to 0, \\
F\left(\mathbf{a}, \mathbf{v}_{ik}\right) - F\left(\mathbf{a}_k, \mathbf{v}_{ik}\right) \to 0, \qquad i = 2, \ldots, n,
\end{gathered}$$

it remains to show that $|\mathbf{v}_{ik}|$ is bounded for $i = 2, \ldots, n$. But this follows from the fact that

$$F\left(\mathbf{a}_k, \mathbf{v}_{1k}\right) + \cdots + F\left(\mathbf{a}_k, \mathbf{v}_{nk}\right) \leq F\left(\mathbf{a}_k, \mathbf{u}_k\right) \to F\left(\mathbf{a}, \mathbf{u}\right),$$

because of which

$$F\left(\mathbf{a}_k, \mathbf{v}_{ik}\right) = |\mathbf{v}_{ik}| \, F\left(\mathbf{a}_k, \overline{\mathbf{v}}_{ik}\right) \leq F\left(\mathbf{a}, \mathbf{u}\right) + C,$$

where C is some positive constant. □

Theorem 3.52. *The distance $G(\mathbf{x}, \mathbf{x} + \mathbf{u}s)$ is differentiable at $s = 0+$ for any $(\mathbf{x}, \mathbf{u}) \in \mathbb{T}$, and*

$$\left.\frac{\mathrm{d}G\left(\mathbf{x}, \mathbf{x} + \mathbf{u}s\right)}{\mathrm{d}s+}\right|_{s=0} = \lim_{s \to 0+} \frac{G\left(\mathbf{x}, \mathbf{x} + \mathbf{u}s\right)}{s} = \widehat{F}\left(\mathbf{x}, \mathbf{u}\right).$$

Proof. We prove first that

$$\limsup_{s \to 0+} \frac{G\left(\mathbf{x}, \mathbf{x} + \mathbf{u}s\right)}{s\widehat{F}\left(\mathbf{x}, \mathbf{u}\right)} \leq 1.$$

Let $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ be a minimizing vector chain for $(\mathbf{x}, \mathbf{u})$, so that

$$\widehat{F}\left(\mathbf{x}, \mathbf{u}\right) = F\left(\mathbf{x}, \mathbf{u}_1\right) + \cdots + F\left(\mathbf{x}, \mathbf{u}_n\right).$$

Consider the chain of points

$$\mathbf{x} \; [\mathbf{x} + \mathbf{u}_1 s] \; [\mathbf{x} + (\mathbf{u}_1 + \mathbf{u}_2)\, s] \; \ldots \; [\mathbf{x} + (\mathbf{u}_1 + \cdots + \mathbf{u}_n)\, s],$$

in which the last point coincides with $\mathbf{x} + \mathbf{u}s$. We will generically refer to a point in this chain as

$$\mathbf{x} + (\mathbf{u}_1 + \cdots + \mathbf{u}_i)\, s, \qquad i = 0, 1, \ldots, n,$$

with the obvious convention for $i = 0$. For all sufficiently small s, all these points belong to a compact ball in $\mathfrak{S}$ centered at $\mathbf{x}$. Then, by Theorem 3.31 and the continuity of F, we have, as $s \to 0+$,

$$\frac{D\left[\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_i\right)s\right]\left[\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_{i+1}\right)s\right]}{sF\left(\mathbf{x},\mathbf{u}_{i+1}\right)}$$
$$=\frac{D\left[\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_i\right)s\right]\left[\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_{i+1}\right)s\right]}{F\left(\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_i\right)s,\mathbf{u}_{i+1}s\right)}$$
$$\times\frac{sF\left(\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_i\right)s,\mathbf{u}_{i+1}\right)}{sF\left(\mathbf{x},\mathbf{u}_{i+1}\right)}\to 1,$$

whence

$$\frac{D\mathbf{x}\left[\mathbf{x}+\mathbf{u}_1s\right]\ldots\left[\mathbf{x}+\mathbf{u}s\right]}{s\widehat{F}\left(\mathbf{x},\mathbf{u}\right)}$$
$$=\frac{\sum_{i=0}^{n-1}D\left[\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_i\right)s\right]\left[\mathbf{x}+\left(\mathbf{u}_1+\cdots+\mathbf{u}_{i+1}\right)s\right]}{s\sum_{i=1}^{n}F\left(\mathbf{x},\mathbf{u}_i\right)}\to 1.$$

But then

$$\limsup_{s\to 0+}\frac{G\left(\mathbf{x},\mathbf{x}+\mathbf{u}s\right)}{s\widehat{F}\left(\mathbf{x},\mathbf{u}\right)}=\limsup_{s\to 0+}\frac{G\left(\mathbf{x},\mathbf{x}+\mathbf{u}s\right)}{D\mathbf{x}\left[\mathbf{x}+\mathbf{u}_1s\right]\ldots\left[\mathbf{x}+\mathbf{u}s\right]}\le 1,$$

by the definition of G. We prove next that

$$\liminf_{s\to 0+}\frac{G\left(\mathbf{x},\mathbf{x}+\mathbf{u}s\right)}{s\widehat{F}\left(\mathbf{x},\mathbf{u}\right)}\ge 1.$$

Consider a sequence of chains

$$\mathbf{x}\ \left[\mathbf{x}+\mathbf{v}_{1k}s_k\right]\ \left[\mathbf{x}+\left(\mathbf{v}_{1k}+\mathbf{v}_{2k}\right)s_k\right]\ldots\left[\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{m_kk}\right)s_k\right],\quad k=1,2,\ldots,$$

such that

$$s_k\to 0+,$$

$$\mathbf{v}_{1k}+\cdots+\mathbf{v}_{m_kk}=\mathbf{u},\qquad k=1,2,\ldots,$$

and

$$\frac{D\mathbf{x}\left[\mathbf{x}+\mathbf{v}_{1k}s_k\right]\ldots\left[\mathbf{x}+\mathbf{u}s_k\right]}{G\left(\mathbf{x},\mathbf{x}+\mathbf{u}s\right)}\to 1.$$

Again, it is easy to see that for all k sufficiently large (i.e., s_k sufficiently small) all these chains fall within a compact ball in $\mathfrak{S}$ centered at $\mathbf{x}$. Then, for $i=0,1,\ldots,m_k-1$, by Theorem 3.31 and the continuity of F, as $k\to\infty$,

$$\frac{D\left[\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{ik}\right)s_k\right]\left[\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{i+1,k}\right)s_k\right]}{s_kF\left(\mathbf{x},\mathbf{v}_{i+1,k}\right)}$$
$$=\frac{D\left[\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{ik}\right)s_k\right]\left[\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{i+1,k}\right)s_k\right]}{F\left(\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{ik}\right)s_k,\mathbf{v}_{i+1,k}s_k\right)}$$
$$\times\frac{s_kF\left(\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{ik}\right)s_k,\mathbf{v}_{i+1,k}\right)}{s_kF\left(\mathbf{x},\mathbf{v}_{i+1,k}\right)}\to 1$$

uniformly across all choices of $\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{m_k k}\right)$. It follows that

$$\frac{D\mathbf{x}\left[\mathbf{x}+\mathbf{v}_{1k}s_k\right]\ldots\left[\mathbf{x}+\mathbf{u}s_k\right]}{s_k\sum_{i=1}^{m_k}F\left(\mathbf{x},\mathbf{v}_{ik}\right)}$$
$$=\frac{\sum_{i=0}^{m_k-1}D\left[\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{ik}\right)s_k\right]\left[\mathbf{x}+\left(\mathbf{v}_{1k}+\cdots+\mathbf{v}_{i+1,k}\right)s_k\right]}{s_k\sum_{i=1}^{m_k}F\left(\mathbf{x},\mathbf{v}_{ik}\right)}\to 1.$$

But then

$$\liminf_{s\to 0+}\frac{G\left(\mathbf{x},\mathbf{x}+\mathbf{u}s\right)}{s\widehat{F}\left(\mathbf{x},\mathbf{u}\right)}=\liminf_{k\to\infty}\frac{D\mathbf{x}\left[\mathbf{x}+\mathbf{v}_{1k}s_k\right]\ldots\left[\mathbf{x}+\mathbf{u}s_k\right]}{s_k\widehat{F}\left(\mathbf{x},\mathbf{u}\right)}$$
$$=\liminf_{k\to\infty}\frac{\sum_{i=1}^{m_k}F\left(\mathbf{x},\mathbf{v}_{ik}\right)}{\widehat{F}\left(\mathbf{x},\mathbf{u}\right)}\geq 1,$$

by the definition of $\widehat{F}$ in terms of minimizing chains. This establishes

$$\lim_{s\to 0+}\frac{G\left(\mathbf{x},\mathbf{x}+\mathbf{u}s\right)}{s\widehat{F}\left(\mathbf{x},\mathbf{u}\right)}=1,$$

and the theorem is proved. □

Theorem 3.56. *For every path* $\mathbf{h}|\left[a,b\right]$ *connecting* $\mathbf{a}$ *to* $\mathbf{b}$ *one can find a piecewise linear path from* $\mathbf{a}$ *to* $\mathbf{b}$ *which is arbitrarily close to* $\mathbf{h}|\left[a,b\right]$ *pointwise and in its length.*

Proof. Let

$$\mu_n=\left\{a=t_{n0},\ldots,t_{ni},t_{n,i+1},\ldots,t_{n,k_n+1}=b\right\}$$

be a sequence of nets with $\delta\mu_n\to 0$. Since the set $\mathbf{h}\left(\left[a,b\right]\right)$ is compact, n can be chosen sufficiently large so that any two successive $\mathbf{h}\left(\alpha=t_{ni}\right)$ and $\mathbf{h}\left(\beta=t_{n,i+1}\right)$ can be connected by a straight-line segment

$$\mathbf{s}_{ni}\left(t\right)=\mathbf{h}\left(\alpha\right)+\frac{\mathbf{h}\left(\beta\right)-\mathbf{h}\left(\alpha\right)}{\beta-\alpha}\left(t-\alpha\right).$$

Then n can further be increased to ensure

$$1-\varepsilon<\frac{G\mathbf{h}\left(\alpha\right)\mathbf{h}\left(\beta\right)}{\widehat{F}\left(\mathbf{h}\left(\alpha\right),\mathbf{h}\left(\beta\right)-\mathbf{h}\left(\alpha\right)\right)}<1+\varepsilon$$

and

$$1-\varepsilon<\frac{D\mathbf{s}_{ni}|\left[\alpha,\beta\right]}{\widehat{F}\left(\mathbf{h}\left(\alpha\right),\mathbf{h}\left(\beta\right)-\mathbf{h}\left(\alpha\right)\right)}<1+\varepsilon.$$

The latter follows from

$$D\mathbf{s}_{ni}|\left[\alpha,\beta\right]=\int_{\alpha}^{\beta}\widehat{F}\left(\mathbf{h}\left(x\right),\dot{\mathbf{h}}\left(x\right)\right)dx=\widehat{F}\left(\mathbf{h}\left(\xi\right),\frac{\mathbf{h}\left(\beta\right)-\mathbf{h}\left(\alpha\right)}{\beta-\alpha}\right)\left(\beta-\alpha\right),$$

for some $\alpha \leq \xi \leq \beta$. Combining the two double-inequalities, for any $\delta > 0$ and all sufficiently large n,

$$1 - \delta < \frac{G\mathbf{h}\left(t_{ni}\right)\mathbf{h}\left(t_{n,i+1}\right)}{D\mathbf{s}_{ni}\left|\left[t_{ni}, t_{n,i+1}\right]\right.} < 1 + \delta,$$

whence

$$1 - \delta < \frac{\sum_{i=0}^{k_n} G\mathbf{h}\left(t_{ni}\right)\mathbf{h}\left(t_{n,i+1}\right)}{D\mathbf{s}_n\left|\left[a,b\right]\right.} < 1 + \delta,$$

where $\mathbf{s}_n\left|\left[a,b\right]\right.$ is the *piecewise linear* path concatenating together $\mathbf{s}_{ni}\left|\left[t_{ni}, t_{n,i+1}\right]\right.$, $i = 0, \ldots, k_n$. By the definition of $D\mathbf{h}\left|\left[a,b\right]\right.$, we have then

$$\lim_{n\to\infty} D\mathbf{s}_n\left|\left[a,b\right]\right. = D\mathbf{h}\left|\left[a,b\right]\right..$$

Since it is obvious that, as $n \to \infty$, $\mathbf{s}_n\left|\left[a,b\right]\right.$ tends to $\mathbf{h}\left|\left[a,b\right]\right.$ pointwise, the theorem is proved. □

References

Aczél, J. (1987). *A short course on functional equations*. Dordrecht: Springer.

Blumenthal, L. M. (1953). *Theory and applications of distance geometry*. London: Oxford University Press.

Blumenthal, L. M., & Menger, K. (1970). *Studies in geometry*. San Francisco, CA: W.H. Freeman.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer-Verlag.

Busemann, H. (2005). *The geometry of geodesics*. Mineola, NY: Dover.

Colonius, H., & Dzhafarov, E. N. (2012). Ultrametric Fechnerian Scaling of discrete object sets. In M. Deza, M. Petitjean, & K. Markov (Eds.), *The mathematics of distances and applications* (pp. 129–132). Sofia, Bulgaria: ITHEA Publisher.

Creelman, C. D. (1967). Empirical detectability scales without the jnd. *Perceptual and Motor Skills*, *24*, 1079–1084.

Dzhafarov, D. D., & Dzhafarov, E. N. (2011). The equivalence of two ways of computing distances from dissimilarities for arbitrary sets of stimuli. *Journal of Mathematical Psychology*, *55*, 469–472.

Dzhafarov, E. N. (2002). Multidimensional Fechnerian scaling: Pairwise comparisons, regular minimality, and nonconstant self-similarity. *Journal of Mathematical Psychology*, *46*, 583–608.

Dzhafarov, E. N. (2003a). Thurstonian-type representations for "same–different" discriminations: Deterministic decisions and independent images. *Journal of Mathematical Psychology*, *47*, 208–228.

Dzhafarov, E. N. (2003b). Thurstonian-type representations for "same–different" discriminations: Probabilistic decisions and interdependent images. *Journal of Mathematical Psychology*, *47*, 229–243.

Dzhafarov, E. N. (2008a). Dissimilarity cumulation theory in arc-connected spaces. *Journal of Mathematical Psychology*, *52*, 73–92.

Dzhafarov, E. N. (2008b). Dissimilarity cumulation theory in smoothly connected spaces. *Journal of Mathematical Psychology*, *52*, 93–115.

Dzhafarov, E. N. (2009). Corrigendum to: "Dissimilarity cumulation theory in arc-connected spaces." *Journal of Mathematical Psychology*, *53*, 300.

Dzhafarov, E. N. (2010). FSCAMDS-Fechnerian Scaling followed by clustering and MDS. *MATLAB program version*. Retrieved from http://www.psych.purdue.edu/

Dzhafarov, E. N. (2010a). Dissimilarity cumulation as a procedure correcting for violations of triangle inequality. *Journal of Mathematical Psychology*, *54*, 284–287.

Dzhafarov, E. N., & Colonius, H. (1999). Fechnerian metrics in unidimensional and multidimensional stimulus spaces. *Psychonomic Bulletin and Review*, *6*, 239–268.

Dzhafarov, E. N., & Colonius, H. (2001). Multidimensional Fechnerian scaling: Basics. *Journal of Mathematical Psychology*, *45*, 670–719.

Dzhafarov, E. N., & Colonius, H. (2006a). Reconstructing distances among objects from their discriminability. *Psychometrika*, *71*, 365–386.

Dzhafarov, E. N., & Colonius, H. (2006b). Regular minimality: A fundamental law of discrimination. In H. Colonius & E. N. Dzhafarov (Eds.), *Measurement and representation of sensations* (pp. 1–46). Mahwah, NJ: Erlbaum.

Dzhafarov, E. N., & Colonius, H. (2007). Dissimilarity cumulation theory and subjective metrics. *Journal of Mathematical Psychology*, *51*, 290–304.

Dzhafarov, E. N., & Colonius, H. (2011). The Fechnerian idea. *American Journal of Psychology*, *124*(2), 127–140.

Dzhafarov, E. N., & Dzhafarov, D. D. (2010). Sorites without vagueness II: Comparative sorites. *Theoria, 76*, 25–53.

Dzhafarov, E. N., & Paramei, G. V. (2010). Space of facial expressions: Cumulated versus transformed dissimilarities. In A. Bastianelli & G. Vidotto (Eds.), *Fechner Day 2010* (pp. 605–610). Padua, Italy: The International Society for Psychophysics.

Dzhafarov, E. N., & Perry, L. (2014). Perceptual matching and sorites: Experimental study of an ancient Greek paradox. *Attention, Perception, and Psychophysics*, *76*, 2441–2464.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*, 5th ed. New York: John Wiley & Sons.

Falmagne, J. C. (1971). The generalized Fechner problem and discrimination. *Journal of Mathematical Psychology*, *8*, 22–43.

Fechner, G. T. (1860). *Elemente der Psychophysik [Elements of psychophysics]*. Leipzig: Breitkopf & Härtel.

Fechner, G. T. (1877). *In Sachen der Psychophysik. [In the matter of psychophysics]*. Leipzig: Breitkopf & Härtel.

Fechner, G. T. (1887). Über die psychischen Massprinzipien und das Webersche Gesetz [On the principles of mental measurement and Weber's law]. *Philosophische Studien*, *4*, 161–230.

Helmholtz, H. v. (1891). Versuch einer erweiterten Anwendung des Fechnerschen Gesetzes im Farbensystem. [An attempt at a generalized application of Fechner's law to the color system]. *Zeitschrift für die Psychologie und die Physiologie der Sinnesorgane*, *2*, 1–30.

Hocking, J. H., & Young, G. S. (1961). *Topology*. Reading, MA: Addison-Wesley.

Kelly, J. L. (1955). *General topology*. Toronto: Van Nostrand.

Krantz, D. (1971). Integration of just-noticeable differences. *Journal of Mathematical Psychology*, *8*, 591–599.

Kujala, J. V., & Dzhafarov, E. N. (2008). On minima of discrimination functions. *Journal of Mathematical Psychology*, *52*, 116–127.

Kujala, J. V., & Dzhafarov, E. N. (2009a). Regular minimality and Thurstonian-type modeling. *Journal of Mathematical Psychology*, *53*, 486–501.

Kujala, J. V., & Dzhafarov, E. N. (2009b). A new definition of well-behaved discrimination functions. *Journal of Mathematical Psychology*, *53*, 593–599.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, *24*, 178–191.

Luce, R. D., & Edwards, W. (1958). The derivation of subjective scales from just noticeable differences. *Psychological Review*, *65*, 222–237.

Luce, R. D., & Galanter, E. (1963). Discrimination. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 191–244). New York: John Wiley & Sons.

Pfanzagl, J. (1962). Über die stochastische Fundierung des psychophysischen Gesetzes. [On stochastic foundations of the psychophysical law]. *Biometrische Zeitschrift*, *4*, 1–14.

Schrödinger, E. (1920/1970). Outline of a theory of color measurement for daylight vision. In D. L. MacAdam (Ed.), *Sources of color science* (pp. 397–447, 481–520). Cambridge, MA: MIT Press.

Schrödinger, E. (1926/1970). Thresholds of color differences. In D. L. MacAdam (Ed.), *Sources of color science* (pp. 183–193). Cambridge, MA: MIT Press.

Ünlü, A., Kiefer, T., & Dzhafarov, E. N. (2009). Fechnerian scaling in R: The package fechner. *Journal of Statistical Software*, *31*(6), 1–24.

4 Mathematical Models of Human Learning

F. Gregory Ashby, Matthew J. Crossley, and Jeffrey B. Inglis

4.1 Early Models of Human Learning

Many early learning theories were seeded by the seminal work of Thorndike. In his famous "puzzle box" experiments, Thorndike placed an animal inside a box with a door that could be opened via a latch accessible to the animal. When the animal learned to operate the latch correctly, the door opened, and it was free to consume a reward placed near the box. Thorndike measured the amount of time it took animals to solve such puzzle boxes and found that the escape time tended to decrease with each trial – that is, the animals learned. From

these observations, Thorndike (1927) postulated the *law of effect*, which states that behavior is driven by associations between stimuli and responses, and that these associations are strengthened when a response is followed by a satisfying effect and weakened when followed by a discomforting effect. With this, the field of associative learning was born. Already apparent in this early work are its clear connections to modern-day reinforcement learning (RL) theory.

Russian physiologist Pavlov (1927) pioneered one still-modern approach to studying associative learning called *classical conditioning*. His famous experiments studying how the salivation response of dogs could be conditioned to occur to a previously neutral stimulus gave the field a standardized paradigm and a new nomenclature (e.g., unconditioned stimulus [US], unconditioned response [UR], conditioned stimulus [CS], and conditioned response [CR]) that drove research in the field forward. Later, Skinner (1938) pioneered many more of the standard methods in use today for the investigation of associative learning. He created operant conditioning chambers – popularly known as the Skinner box – that were equipped with both a manipulandum (e.g., a lever) and a tool to record lever pulls so that cumulative operant behavior (e.g., pulling the lever) could be measured over an experimental session. This approach came to be known as *operant* or *instrumental conditioning*.

Watson and Guthrie followed many of the basic tenets of associative learning formulated by Thorndike, but each introduced novel refinements (e.g., Guthrie, 1935; Watson, 1913). Unlike Thorndike, neither thought that reinforcement (i.e., neither a satisfying nor a discomforting effect) was necessary for associative learning. Rather, they thought that mere temporal contiguity between stimulus and response was sufficient. Later in this chapter, we will see how this notion is related to a form of two-factor synaptic plasticity proposed by Hebb (1949).

An important theoretical alternative to the dominant theories of instrumental conditioning came from Tolman (1948), who advocated that animals learned "cognitive maps" and used these maps to make flexible and goal-directed actions. This view gained relatively little traction in Tolman's lifetime, but is renewed today by modern model-based RL accounts of learning.

The first attempts to formalize theories of learning focused on building mathematical equations that could fit learning curves from a variety of different conditioning experiments (Gulliksen, 1934; Hull, 1943; Thurstone, 1919). The most systematic attempts were by Hull (1943), who embraced Thorndike's fundamental ideas on associative learning – although he spoke of *habits* instead of stimulus–response associations. More importantly, he expressed his views in the form of explicitly stated assumptions. The resulting equations clearly expressed what Hull believed were driving factors of an animal's behavior (e.g., habit strength, drive reduction, etc.).

In the 1950s, mathematical models of learning began to focus less on curve fitting and more on the psychological processes that mediate the learning. This change in focus began with two *Psychological Review* articles on mathematical learning theory that appeared in quick succession – Estes' (1950) introduction of stimulus-sampling theory and Bush and Mosteller's (1951) description of the

linear-operator model. Both of these contributions were hugely influential, partly because they were among the first process models in psychology and, as such, they spurred others to develop their own process models. The excitement created by these efforts played a key role in the birth of modern mathematical psychology. But both articles were also influential in their own right. In particular, the linear-operator model inspired the Rescorla–Wagner model (Rescorla & Wagner, 1972), which is now ubiquitous in the learning literature, and more than a half century later, stimulus-sampling theory continues to motivate new research (e.g., Fanselow *et al.*, 2014; Soto & Wasserman, 2010).

Mathematical learning theory played a huge role in the field of mathematical psychology during its first formal decade of existence. Stimulus-sampling theory and the linear-operator model were both elaborated, and a large number of Markov chain models were proposed that assumed learning was a process of moving between discrete states of knowledge. During the 1960s, interest in cognitive processes saw a shift to models of concept learning, which today would be called rule-based learning, and a new focus on the cognitive components of learning, including attention, storage, and retrieval (e.g., Greeno & Bjork, 1973). Much of this work is reviewed in the classic text by Atkinson, Bower, and Crothers (1965).

Today, mathematical models of learning are developed and tested in a wide range of different fields, including, for example, machine learning (e.g., Alpaydin, 2020; Mohri, Rostamizadeh, & Talwalkar, 2018) and learning in simple species such as *Drosophila* (e.g., Kennedy, 2019) and zebrafish (e.g., Ninkovic & Bally-Cuif, 2006). A review of all this work is outside the scope of any one chapter. Instead, our focus will be on mathematical models of human learning. In some cases, we will consider developments in machine learning and research with nonhuman animals, but in all cases the focus will be on how such work has contributed to our understanding of human learning.

4.2 Neuroscience Breakthroughs

Mathematical modeling of human learning began to languish in the late 1960s, partly because of the cognitive revolution that turned interest to other phenomena, and partly because it became apparent that the best existing models were valid for only a narrow and limited set of learning-related phenomena. Furthermore, models that succeeded in different domains often bore little similarity to each other. This landscape remained largely unchanged for the next several decades, until two breakthroughs in neuroscience offered a clear path forward. The first was the discovery of long-term potentiation and long-term depression, which served as promising models of learning at the cellular level. The second breakthrough was the discovery that humans have multiple learning and memory systems that for the most part are functionally and anatomically distinct, and that each control behavior under different experimental conditions. As a result, it is likely that no single mathematical model can describe all human learning. Instead, qualitatively different models are needed for different learning systems.

4.2.1 Discovery of LTP and LTD

In his classic 1949 book entitled *Organization of Behavior: A Neuropsychological Theory*, Donald Hebb proposed a neural mechanism that he thought might mediate learning and memory. Specifically, he postulated:

> Let us assume then that the persistence or repetition of a reverberatory activity (or 'trace') tends to induce lasting cellular changes that add to its stability. The assumption can be precisely stated as follows: When an axon of cell A is near enough to excite a cell B and repeatedly and persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased. (Hebb, 1949, p. 62)

Hebb's hypothesis is now widely known as *Hebbian learning*.

Several decades later, this exact type of neural plasticity was discovered at synapses in the hippocampus (Bliss & Lømo, 1973). Specifically, brief, high-frequency presynaptic activation was found to cause a persistent (at least 1 hour) increase in the postsynaptic response – a phenomenon known as *long-term potentiation* (LTP). Then, 9 years later, the opposite phenomenon of *long-term depression* (LTD) was discovered, in which prolonged, but weak, presynaptic activation causes a persistent (at least 1 hour) decrease in the postsynaptic response (Ito, Sakurai, & Tongroach, 1982). LTP and LTD have now been observed and closely studied in many different brain regions and in many different cell types. Furthermore, they are known to occur under a plethora of diverse conditions, and to be driven by numerous intracellular signalling cascade mechanisms. Although a review of the current literature on LTP and LTD is well beyond the scope of this chapter, a noncontroversial conclusion of this literature is that it is now widely accepted that LTP and LTD form the neural basis of learning and memory (e.g., Martin, Grimwood, & Morris, 2000; Nicoll, 2017).

4.2.2 Discovery of Multiple Learning and Memory Systems

Early mathematical models of learning assumed that all human learning occurs in the same way, which suggests that all learning should depend on the same neural network and be consolidated into the same memory system. This assumption was inconsistent with the growing body of evidence that began to accumulate in the 1960s showing that the best models seemed valid for only a narrow range of experimental tasks, and this led many mathematical psychologists to turn away from the study of learning. A resurgence in mathematical models of learning was ushered in by the discovery that humans have multiple learning and memory systems that for the most part are functionally and anatomically distinct, that evolved at different times and for different purposes, that are ideally suited to learning different types of information, and that thrive under very different environmental conditions.

The first step in this process was to realize that humans have multiple memory systems (e.g., Eichenbaum & Cohen, 2001; Poldrack *et al.*, 2001; Squire, 2004;

Tulving & Craik, 2000). After overwhelming evidence in support of multiple memory systems was documented, it was an easy inference to conclude that humans must therefore also have multiple learning systems. After all, learning is the acquisition of a skill or some form of knowledge, and memory is the storage and/or expression of what was learned. So learning and memory are closely related. Mathematical models of learning focus on how the memory traces are established and consolidated, whereas models of memory focus on the nature of those traces and how they are accessed to produce memory-dependent behaviors (e.g., see Chapter 5 in this volume). For this reason, an obvious hypothesis is that there are as many learning systems as there are memory systems (e.g., Ashby & Maddox, 2005; Ashby & O'Brien, 2005).

As soon as the multiple-systems hypothesis was formulated, work began to identify the networks that mediate learning in each system and to study the properties of the various systems (for a review, see e.g., Ashby & Valentin, 2017). This body of research made it clear that no single model was likely to account for all human learning. For example, basal-ganglia-mediated procedural learning is incremental, whereas prefrontal-cortex-mediated rule learning is mostly all-or-none (e.g., J. D. Smith & Ell, 2015).

4.3 Modern Approaches to Modeling Human Learning

The birth of mathematical psychology coincided with the first attempts to build process models of learning. The reinterest in learning that occurred with the neuroscience breakthroughs described in the previous section coincided with the development of new types of learning models, and also with the first ever implementational-level models – that is, models that attempt to describe the neural circuits that implement the algorithms described by process models. This section briefly introduces these more modern approaches to building mathematical models of learning, and then the rest of the chapter examines these trends in more detail.

4.3.1 Descriptive- and Process-Level Approaches

Current descriptive and process models of human learning are dominated by two different, but converging, approaches – one rooted in the statistics literature and one rooted in the machine-learning and computer-science literatures (as described, e.g., by Alpaydin, 2020; Sutton & Barto, 1998). Both attempt to build models that optimize some aspect of learning – the former by following principles of Bayesian statistics, and the latter by assuming that human learning depends on some popular machine-learning algorithms.

Normative models have a long history in psychology. For example, ideal observer models have played an important role in psychophysics and signal detection theory since the 1950s (e.g., Green & Swets, 1966). Similarly, during the 1980s and 1990s, human classification performance was carefully compared to the performance of optimal classifiers (e.g., Ashby & Alfonso-Reese, 1995;

Ashby & Maddox, 1998). Comparing human performance to the performance of an optimal device is a valuable step in the evolution of model building in any area of psychology. Humans are highly skilled in many behaviors, so an optimal model will often provide a reasonably good fit to human data. Better fits are usually possible by adding certain suboptimal components to the model, such as various types of noise. Carefully documenting which types of added suboptimalities allow the model to provide the best fit provides invaluable information about the underlying processes that mediate the behavior. In the case of human learning, the Bayesian models are objectively optimal, in the sense that they assume the learner chooses the response most likely to be correct, and that these choice probabilities are updated trial-by-trial according to Bayes' theorem. In this class of models, learning is typically equivalent to Bayesian updating.

An alternative approach, which is perhaps even more popular and looks very different on the surface, is to build models that assume human learning follows algorithms that were developed in the computer-science, machine-learning, and artificial-intelligence literatures. In this approach, the models are typically some form of neural network, and learning is a process of adjusting the connection strengths or weights between units. These algorithms fall into one of three general classes: unsupervised, RL, and supervised (e.g., Alpaydin, 2020). Unsupervised learning algorithms, which include Hebbian learning as a prominent special case, modify all learning-related weights using the same algorithm and without regard to feedback. RL algorithms also apply the same learning algorithm to every weight, but the algorithm applied depends on the type of feedback that was delivered (e.g., reward vs. non reward). Finally, supervised learning algorithms attempt to compute the unique error of the output unit associated with every modifiable weight in the network, and they then tune that weight according to this unique error. The most prominent examples, such as backpropagation and the delta rule, attempt to implement a gradient descent optimization procedure. Unsupervised learning and RL are *global learning rules* because they apply the same rule to every learning-related weight, whereas supervised learning is a *local learning rule* because it uses a different error to modify every weight.

Early models imported from computer science assumed that learning followed gradient descent trajectories, as implemented, for example, by the delta rule and backpropagation. More recently, a large subset of these models apply one of the many RL algorithms that are described in the influential text of Sutton and Barto (1998). Included in this list are temporal-difference learning, actor–critic architectures, Q learning, and SARSA (State–Action–Reward–State–Action).

The models in this class are not objectively optimal, at least not in the sense of the Bayesian models, which try to maximize response accuracy. Even so, the learning algorithms they assume were all developed in attempts to maximize the learning abilities of some artificial system. Therefore, if not objectively optimal, many of them are among the most efficacious learning algorithms ever invented. In this sense, models in this class are similar to the normative models that are constructed using Bayesian statistics.

4.3.2 Implementational-Level Approaches

Implementational-level models require extensive knowledge about brain function and behavior. Despite this high standard, they date back at least to early work by Marr (e.g., Marr, 1969) and Grossberg (e.g., Grossberg, 1972). One remarkable aspect of these early models is that they predate the discovery of the forms of synaptic plasticity that they postulated. Despite this early and seminal work, until recently, there were relatively few implementational-level models in psychology.

During the past two decades, the field of neuroscience has exploded, and the number of implementational-level models in psychology has grown commensurately. As these models became more popular, new methods were developed to build and test them, and collectively this new field is known as *computational cognitive neuroscience* (Ashby, 2018; O'Reilly *et al.*, 2012). The goal here is to first identify the neural network that mediates the behavior and then build a model that mimics neural activity in this network. In the case of learning, the model should display learning-related synaptic plasticity in accord with what is observed in the biological system being modeled. Such models are generally more computationally intractable than their process counterparts, and therefore require extensive computer simulation to fit and test. Even so, despite this cost, implementational models have many advantages over more traditional process models (e.g., see Ashby, 2018). For example, whereas process models can generally be tested only against response time and accuracy data, computational cognitive neuroscience models can be tested against virtually any dependent measure between behavior at the highest level and single-unit recordings at the lowest level, including, for example, response times, accuracies, single-neuron recordings, fMRI blood oxygen-level dependent (BOLD) responses, and EEG recordings. Another advantage is that if two computational cognitive neuroscience models are built and validated that each account for different types of behaviors, then because each should be faithful to the underlying neuroanatomy, it should be possible to link the two in a plug-and-play fashion to create a new composite model that is consistent with all the behavioral and neuroscience data that are consistent with either model alone (as done, e.g., by Cantwell *et al.*, 2017).

Implementational models attempt to model activity in the actual neural circuits that mediate the behavior under study. And rather than borrow learning algorithms from Bayesian statistics or machine learning, they directly model the types of synaptic plasticity thought to occur during LTP and LTD. Thus, whereas implementational models directly model the structures and processes thought to mediate learning, Bayesian models and models based on machine-learning notions of RL are examples of modeling by analogy – in the sense that they are based on algorithms developed for other purposes (statistics or machine learning). Modeling by analogy has a long history in psychology ('the brain is like a telephone switchboard'; 'the brain is like a computer'), and comparing human behavior to other devices can be a useful exercise because it can expose uniquely human characteristics. Implementational models should be the ultimate goal, but they

require far more knowledge to build, and for many behavioral phenomena, this high threshold has not yet been reached.

Whereas it was always obvious that 'the brain is like a telephone switchboard' is an analogy, as the analogies became more sophisticated, they also became more difficult to recognize. This is especially true with models based on machine-learning RL algorithms. After all, RL has been a central focus of research within psychology for more than a century. Furthermore, it was quickly noted that synaptic plasticity in the striatum has properties that are similar to several popular machine-learning RL algorithms (e.g., Doya, 2000; Houk, Adams, & Barto, 1995). Because of this similarity, learning models based on machine-learning notions of RL can be especially useful. Ultimately, though, we should expect that synaptic plasticity, and therefore learning, will have some uniquely human properties that require their own uniquely human models to capture completely.

4.4 Descriptive and Process Models of Human Learning

4.4.1 Reinforcement Learning

In computer science, RL is a general approach to building decision-making agents that learn to maximize rewards. The standard approach (Sutton & Barto, 1998) is to model the environment as a Markov decision process and to assume that the agent moves through a set of discrete states $S = \{s_1, s_2, \ldots, s_n\}$ by choosing among a set of possible actions $A = \{a_1, a_2, \ldots, a_m\}$. The decision rule that determines the probability of each possible action, given a particular state, is called the action policy π.

A state can be almost anything. The one requirement is that since we assume a Markov decision process, the states, as defined by the model, must satisfy the Markov property, in the sense that knowledge of the current state alone should be enough to compute the predicted probability of reward and this probability should not depend on the path the agent took to reach the current state. So, for example, a state could be the position of a rat in a maze. If the rat is in an arm that is not baited with reward, then the probability of imminent reward is low, whereas if the animal is in a baited arm, then the probability of imminent reward is high.

The action policy π determines the probability that the state will change from s_i to s_k, for any i and k. Since each state has some true probability of imminent or future reward that is determined by the environment, the actions selected by the agent therefore also determine current and future reward probabilities. Thus, the agent must learn to take actions that cause transitions to the most rewarding states. This requires that the agent learns the value of each state. Value is formalized in the state-value function $V_\pi(s)$, which equals the expected value of all rewards – both current and future – that the agent can expect if the state is s and future actions are selected according to policy π.[1] Let r_t denote the value of a reward received

1 Sutton and Barto (1998) define the value function as the expected value of all future rewards. Therefore, in their formulation, the current reward does not contribute to the value function. This definition implies that the value to an animal of reaching a baited goal box when exploring a maze is

t time units in the future, and let R denote the total value of all current and future rewards. Then RL models assume

$$R = \sum_{t=0}^{\infty} \gamma^t r_t, \tag{4.1}$$

where $0 < \gamma \leq 1$ is a temporal discounting parameter that serves to reduce the value of rewards the more distant they occur in the future. The value function is then defined as

$$\begin{aligned} V_\pi(s) &= \mathrm{E}[R|\pi, s] \\ &= \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t|\pi, s\right]. \end{aligned} \tag{4.2}$$

Different methods use the estimated value function in different ways to select the best actions, and a complete description of all these methods is beyond the scope of this chapter. However, one major dimension on which different methods are classified is whether or not the agent directly estimates the state transition probabilities (i.e., the probability that the state will transition from s_i to s_k when action a_j is selected). Methods that estimate these transition probabilities are called *model-based*, whereas methods that do not are called *model-free*.

Model-Free RL Approaches

The iterative sample mean. As we have seen, the goal of many RL models and algorithms is to estimate a state-value function. For example, the Rescorla–Wagner model estimates the expected reward value of a cue in a classical conditioning paradigm, temporal-difference learning estimates the expected value of all future rewards given some fixed action policy, and Q learning estimates a similar value for different state–action pairs. The standard statistical approach to parameter estimation assumes a sample of fixed size. RL algorithms, however, apply to an agent operating in real time through an environment that presents successive opportunities to receive rewards. Therefore, the agent must continually update value estimates when moving through the environment. For this reason, parameter estimation must be iterative (e.g., as in dynamic programming). This is a straightforward and well-known statistical problem. For example, a population mean can be estimated iteratively as follows.

Theorem 4.1. *Consider a set of successive samples $X_1, X_2, \ldots, X_n$ that are all drawn from some population. Then the sample mean equals*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_{n-1} + \frac{1}{n}(X_n - \overline{X}_{n-1}), \tag{4.3}$$

where $\overline{X}_0 = 0$. Furthermore, note that X_n is the current sample and $\overline{X}_{n-1}$ is the best guess of X_n after $n - 1$ samples have been collected (i.e., in the sense of the

zero. For this reason, we chose to define the value function as the expected value of all current and future rewards.

law of large numbers). As a result, $X_n - \overline{X}_{n-1}$ is the prediction error – call it PE. So Equation (4.3) is equivalent to

$$\overline{X}_n = \overline{X}_{n-1} + \frac{1}{n} PE. \tag{4.4}$$

Proof. See, for example, Ashby (2018). □

In other words, the standard, batch estimate of the population mean, $\overline{X}_n$, can be efficiently computed in real time by updating the old estimate by an amount that is proportional to the prediction error. If the newest sample is larger than expected (i.e., if $X_n > \overline{X}_{n-1}$) then the mean estimate is increased, and if the newest sample is smaller than expected (i.e., if $X_n < \overline{X}_{n-1}$) then the mean estimate is decreased.

As we will see, the most popular RL algorithms are all based on Equation (4.4). They differ mainly in how they define X_n, although in all RL algorithms the goal is to estimate some reward-related value. In such cases, the prediction error in Equation (4.4) becomes a *reward prediction error* (RPE), which in general is defined as obtained reward minus expected reward.

Because the iterative estimate of the mean is mathematically equivalent to the standard, batch estimate, it possesses the same statistical properties. Therefore, note that this iterative estimate is the uniformly minimum variance unbiased estimator of the population mean if the X_i are independent and identically distributed (i.i.d.) samples from some population, and the sample size n is known ahead of time. In many real-world environments, of course, the samples are not i.i.d., and if the sampling is done in real time, the final sample size is often unknown. The standard RL solution to these problems is to replace the constant $1/n$ with some constant α that can be adjusted or set in a way that depends on the environment. For example, a standard approach is to set α in a way that causes temporal discounting, so that recent samples are weighted more heavily than early samples.[2] In fact, this form of temporal discounting occurs whenever $\alpha > 1/n$. Therefore, when applied to nonstationary data, the *iterative sample mean* is

$$\overline{X}_n = \overline{X}_{n-1} + \alpha(X_n - \overline{X}_{n-1}) = \overline{X}_{n-1} + \alpha PE. \tag{4.5}$$

The parameter α is commonly referred to as the learning rate because increases in α cause $\overline{X}_n$ to change more quickly.

Another advantage of the iterative sample mean, relative to the batch estimate, is that it is easier to incorporate prior beliefs into the estimate of the population mean. For example, consider a simple coin-tossing experiment in which the goal is to estimate the probability of a heads [i.e., where we assign a value of 1 to each heads and 0 to each tails, and then use Equations (4.5) to estimate the true probability of a heads]. A natural prior belief might be that the coin is fair, which is easily incorporated into Equation (4.5) by setting $\overline{X}_0 = 0.5$.

2 Note that we have now introduced two different temporal discounting parameters. The parameter γ discounts future rewards and the parameter α discounts distant samples.

Temporal-difference learning. Temporal-difference learning estimates the state-value function under the assumption that the action policy is fixed. A popular paradigm that satisfies this constraint is classical conditioning, in which some cue may or may not be followed some time later by a reward or perhaps by multiple rewards. The goal of the agent in this case, is to learn that the cue predicts a future reward. Note that there is no action to produce here and so in this special case, we can omit the subscript π in our notation. And although temporal-difference learning applications to classical conditioning are free to define the states in any way that satisfies the Markov property, the most common definition, by far, is to define the states as time points that begin with the cue and end with the last possible reward.

Therefore, define the state $s_t = t$, where t equals the number of time steps since cue presentation, and let $r_n(t)$ equal the value of the reward received at time t on trial n. Then the total value of all future rewards on trial n at time t equals

$$R_n(t) = \sum_{i=t}^{T} \gamma^{i-t} r_n(i), \tag{4.6}$$

where T equals the time of the last possible reward on each trial. As in other RL algorithms, the goal of temporal-difference learning is to estimate the state-value function – that is, the expected value of $R_n(t)$:

$$V_n(t) = \mathrm{E}[R_n(t)]. \tag{4.7}$$

Now, because $V_n(t)$ is a population mean, our best estimate is the sample mean of the $R_i(t)$ across previous trials:

$$\hat{V}_n(t) = \frac{1}{n} \sum_{j=1}^{n} R_j(t). \tag{4.8}$$

This sample mean can be estimated efficiently via term-by-term substitution into the iterative sample mean defined in Equation (4.5) to produce

$$\hat{V}_n(t) = \hat{V}_{n-1}(t) + \alpha[R_n(t) - \hat{V}_{n-1}(t)]. \tag{4.9}$$

The problem with this estimate is that $R_n(t)$ includes the immediate reward $r_n(t)$, plus all future rewards that will be obtained on trial n – that is

$$\begin{aligned} R_n(t) &= r_n(t) + \sum_{i=t+1}^{T} \gamma^{i-t} r_n(i) \\ &= r_n(t) + \gamma \sum_{i=t+1}^{T} \gamma^{i-(t+1)} r_n(i) \end{aligned} \tag{4.10}$$

and unfortunately, all reward-related terms on the right except $r_n(t)$ are unknowable since they occur in the future. Temporal-difference learning estimates the unknowable part – that is, the expression defined by the summation sign – by using

the iterative sample mean of all rewards that occurred after time t on previous trials [i.e., via $\hat{V}_{n-1}(t+1)$]. This results in the following estimate:

$$\hat{R}_n(t) = r_n(t) + \gamma \hat{V}_{n-1}(t+1). \tag{4.11}$$

Substituting this estimate into Equation (4.9) for $R_n(t)$ produces the final form of temporal-difference learning:

$$\hat{V}_n(t) = \hat{V}_{n-1}(t) + \alpha[r_n(t) + \gamma \hat{V}_{n-1}(t+1) - \hat{V}_{n-1}(t)]. \tag{4.12}$$

Note that, despite initial appearances, the expression in square brackets equals the prediction error (or more specifically, the RPE), just as in Equation (4.5). The first term in square brackets is the immediately obtained reward and the second term is the best guess of the (discounted) value of all future rewards expected on trial n. The sum of the first two terms is therefore the agent's estimate of the total obtained rewards on trial n given that we are t time units into the trial. The last term is the predicted value of this quantity that was made before the trial began. Therefore, the sum of the first two terms represents obtained reward, whereas the last term represents predicted reward.

As an application of temporal-difference learning, consider a simple classical-conditioning task in which the same CS (e.g., a light or tone) is followed T time steps later by a reward. On the first presentation of the CS, the subsequent reward is unexpected, but as the CS–reward pairing is repeated, the agent will eventually learn that the CS is paired with future reward. The standard temporal-difference learning application to this task assumes that initially, all states have zero value [i.e., $V_0(t) = 0$, for all t] because the CS has never before been paired with reward. On trial 1, the CS is presented and then the agent unexpectedly receives a reward at time T. Suppose the value of this reward is r. Then temporal-difference learning predicts that

$$\hat{V}_1(T) = \hat{V}_0(T) + \alpha[r_1(T) + \gamma \hat{V}_0(T+1) - \hat{V}_0(T)]. \tag{4.13}$$

Note that, by our assumptions about initial conditions, all V_0 terms equal 0. However, $r_1(T) = r$ because the agent receives a reward on each trial at time T. Therefore

$$\hat{V}_1(T) = \alpha r. \tag{4.14}$$

Now consider the value that temporal-difference learning assigns to the state that is one time unit earlier than reward delivery on trial 2:

$$\hat{V}_2(T-1) = \hat{V}_1(T-1) + \alpha[r_2(T-1) + \gamma \hat{V}_1(T) - \hat{V}_1(T-1)]. \tag{4.15}$$

Note that $\hat{V}_1(T-1) = 0$ because at time $T-1$ of trial 1 the agent has not yet received any rewards. Furthermore, $r_2(T-1) = 0$ because rewards are delivered at time T, not at time $T-1$. However, as we saw, $\hat{V}_1(T) = \alpha r$. Therefore

$$\hat{V}_2(T-1) = \alpha^2 \gamma r. \tag{4.16}$$

In other words, the RPE that occurred at time T on trial 1 has propagated back on trial 2 to the immediately preceding state (i.e., $T - 1$). Similarly, on trial 3, the positive value associated with state $T - 1$ will propagate back to state $T - 2$. In this way, the value associated with earlier and earlier states will increase. This propagation will continue until it eventually reaches the time of cue presentation – that is, until $V_n(0) > 0$, for some value of n. It will not propagate to earlier times than this, however, so long as cue presentation times are unpredictable.

Temporal-difference learning is popular, in part because it shares some properties with the firing properties of dopamine (DA) neurons. In particular, in this same classical-conditioning experiment, DA neurons will eventually begin to fire to any cue that predicts a future reward. We will consider temporal-difference learning as a model of DA neuron firing in more detail in a later section.

Q-learning. Q-learning is a model-free RL algorithm to learn actions that maximize current and future rewards. It is similar to temporal-difference learning, but it learns the value of state–action pairs, instead of states independently of the selected action. The resulting value function, denoted by $Q_n(s, a)$ (i.e., "Q" for quality), gives the value of taking action a from state s on trial n, under the assumption that all actions after a are optimal with respect to the estimated action-value function – that is, that all future actions are selected so as to maximize total reward. The policy that always chooses the action that maximizes reward is called the *greedy policy*. So Q learning updates the value function under the assumption that a greedy policy will be used, even when the agent follows some nongreedy policy. Algorithms that estimate the value function using a policy that is different from the one that is currently being followed are called *off-policy* algorithms.

Let s_t denote the state at time t and a_t denote the action taken at time t. Let $R_n(a_t|s_t)$ denote the total current and future rewards obtained on trial n if the state is s_t, action a_t is immediately taken, and all subsequent actions are greedy. Then term-by-term substitution into the iterative sample mean produces

$$\hat{Q}_n(s_t, a_t) = \hat{Q}_{n-1}(s_t, a_t) + \alpha[R_n(a_t|s_t) - \hat{Q}_{n-1}(s_t, a_t)]. \tag{4.17}$$

As with temporal-difference learning, R_n is unknowable since it depends on future rewards. So Q learning estimates R_n via

$$\hat{R}_n(a_t|s_t) = r_n(t) + \gamma \max_a \hat{Q}_{n-1}(s_{t+1}, a); \tag{4.18}$$

that is, by adding the immediate reward to the discounted (iterative sample) mean of the best rewards produced by any sequence of past actions that started from the state that results from taking action a_t from state s_t. Combining these two equations produces the final form of Q learning:

$$\hat{Q}_n(s_t, a_t) = \hat{Q}_{n-1}(s_t, a_t) + \alpha \left\{ r_n(t) + \gamma[\max_a \hat{Q}_{n-1}(s_{t+1}, a)] - \hat{Q}_{n-1}(s_t, a_t) \right\}. \tag{4.19}$$

A common assumption is that all initial Q values equal 0 [i.e., $Q_0(s, a) = 0$, for all s and a]. Once these initial values are all set, Equation (4.19) is used to update the Q estimates beginning on trial 1.

The Q values are often used to define action policies. For example, as we already saw, the greedy action policy is to always choose the action with the maximum Q value. Although, at first glance, this policy sounds appealing, note that it fails to explore the set of all possible actions. Unless the optimal policy is discovered early on by chance, then the greedy policy is unlikely to ever discover this optimal policy. Therefore, many action policies trade off exploration and exploitation. One way to do this is via an ϵ-greedy algorithm that selects an action randomly with probability ϵ and uses a greedy policy with probability $1 - \epsilon$. Another popular choice is to compute the action selection probabilities by passing the Q values through a softmax function:

$$P(a_i|s) = \frac{e^{Q(s, a_i)}}{\sum_j e^{Q(s, a_j)}}. \tag{4.20}$$

Note that since this policy depends on the Q values, updating or changing the estimates of Q changes the policy.

To make the discussion concrete, consider an agent in a maze in which one or more arms are baited with reward. In this case, we can consider the states to be locations within the maze, and the actions to be movements that carry the agent from one location to another. Before any learning has occurred, if state s_i is one step before a baited arm, then the action that moves the animal one step forward (i.e., towards the reward) will be rewarded and $Q(s_i, a_{\text{forward}})$ will gain positive value.

Model-Based RL Approaches

Model-based RL approaches build a model of the environment by estimating the value function [e.g., $V_\pi(s)$ or $Q(s, a)$], and the state-transition function $T(s_k|a_j, s_i)$, which specifies the probability that taking action a_j will transition the agent from state s_i to state s_k. Once accurate estimates of these functions are available, action selection is a matter of solving directly for the action sequence that maximizes reward.

Daw, Niv, and Dayan (2005) proposed a dual-controller model that assumes the brain includes both model-free and model-based RL algorithms, with behavior determined by the system that is most confident in its predictions (i.e., has the lowest uncertainty). The model includes a striatal-mediated, model-free cache system that implements habit learning and a prefrontal-cortex-mediated model-based tree-search system that implements goal-directed learning. Both systems use a form of Q learning. Traditional Q learning [i.e., Equation (4.19)] does not track uncertainty, so Daw, Niv, and Dayan (2005) proposed a Bayesian version (i.e., based on Dearden, Friedman, & Russell, 1998), in which both systems attempt to estimate a distribution of Q values across trials. If we assume that rewards are Bernoulli distributed, then a convenient prior distribution of Q values is the beta distribution (because the beta distribution is a conjugate prior for the Bernoulli

distribution). This prior is then updated through Bayes' rule to obtain model-free and model-based posterior distributions of Q values.

The model-based system consists of a Bayesian tree-search algorithm in which the agent uses experience in its environment to estimate the distribution of reward values $R(a_j|s_i)$ and state-transition probabilities $T(s_k|a_j, s_i)$. The model assumes a beta distribution for the prior on rewards and a Dirichlet distribution for the prior on the state transitions. These distributions are then updated according to Bayes' rule by counting up the obtained rewards and state transitions. Since both systems estimate distributions of Q values, the variances from these two estimates are compared and the system with the lowest uncertainty controls the response of the agent.

The model successfully accounts for a variety of instrumental conditioning phenomena, including, for example, the effects of reward devaluation (Dickinson & Balleine, 2002). In these experiments, an animal is trained to lever press (for example) and at some point in training, the reward is devalued prior to the session, typically either by providing free access to food or administering a drug that causes ingestion of the reward to induce illness. Early in training, reward devaluation reduces the frequency of the instrumental behavior, but after extensive overtraining, the behavior becomes immune to the devaluation. Furthermore, the degree to which the animal is sensitive to devaluation is proportional to the complexity of the task and the temporal proximity of the action to the reward. The dual-controller model accounts for these phenomena by proposing that the model-based tree-search system controls responding early in training, but that control is passed to the model-free cache system after overtraining. Furthermore, the amount of training required for the transfer of control is assumed to increase with the complexity of the task.

Early in training, new information immediately influences action values at all states in the tree-search system. In contrast, the cache system takes significantly longer to propagate new information to other states. Additionally, in more complex tasks, the tree-search system takes control because it is more data efficient – in more complex tasks there is less data available for each state–action pair. Finally, the tree-search system performs better for actions more proximate to reward due to its superior data efficiency, whereas the cache system performs better for actions more distal to reward due to its lower sensitivity to computational noise. Since the tree-search system has a model of the task and access to the long-term consequences of its actions at each time step, it can adapt its policy in response to reward devaluation. Alternatively, the cache system estimates the value of each action directly, and reward devaluation is insufficient to reverse the cumulative effects of the many positive rewards that were received earlier in training.

4.4.2 Bayesian Modeling of Human Learning Under Uncertainty

The RL models described in the previous section are arguably more popular than Bayesian models of learning, at least partly because they are computationally

simpler to implement. Traditional Bayesian approaches require numerical evaluation of complex multiple integrals. This section reviews the hierarchical Gaussian filter (Mathys *et al.*, 2011, 2014), which attempts to overcome this limitation by deriving computationally simple, interpretable, and efficient update equations – similar to those used in RL models – except from normative Bayesian principles. Conveniently, these update equations also enable the estimation of agent-specific parameters that allow each individual to be modeled as subjectively optimal with respect to minimizing the agent's surprise (i.e., free energy) when unexpected events occur.[3] Furthermore, the form of these update equations is similar to a version of the iterative sample mean [Equation (4.5)] in which the learning rate, α, is modulated by various forms of uncertainty. Accordingly, the benefits of the hierarchical Gaussian filter extend past the Bayesian framework by providing a normative foundation for the sequential updating equations of heuristic RL algorithms.

As a context for describing the model, consider an (A, not A) categorization task in which an agent is asked to report whether or not a presented stimulus belongs to category A (e.g., by responding YES or NO). Suppose the stimuli in category A vary on one stimulus dimension, call it w, and are normally distributed on this dimension with mean μ_A and variance π_A^{-1}; that is

$$w \sim \mathcal{N}(\mu_A, \pi_A^{-1}), \tag{4.21}$$

where π_A is the precision of the category A samples (not to be confused with action policies as defined in the RL literature). Suppose that on "not A" trials, the stimuli are uniformly distributed on dimension w over all physically realizable values. Therefore, the optimal decision strategy is to respond YES if the presented stimulus is close to μ_A on dimension w and NO if it is far away. Consider the simplest possible case in which the agent knows π_A but not μ_A. Then the optimal strategy requires the agent to estimate μ_A.

An agent trying to estimate μ_A could do so by computing the iterative sample mean [Equation (4.5)]. Instead, however, consider a Bayesian approach. Suppose the agent assumes that the prior distribution of μ_A is

$$\mu_A \sim \mathcal{N}(\mu_0, \pi_0^{-1}), \tag{4.22}$$

where π_0 is the precision of the agent's knowledge of the task. Suppose further that on trial n the stimulus has value w_n on the relevant dimension and the feedback informs the agent that this stimulus belonged to category A. Then a Bayesian approach indicates that the posterior likelihood that the true orientation is μ_A is:

$$p(\mu_A|w_n) = \frac{p(w_n|\mu_A)p(\mu_A)}{\int p(w_n|\mu)p(\mu)d\mu} \sim \mathcal{N}\left(\mu_{\mu_{A|w_n}}, \pi^{-1}_{\mu_{A|w_n}}\right). \tag{4.23}$$

3 See Ashby (2019), Friston *et al.* (2007), or Penny (2012) for a description of free-energy minimization in the context of model selection, and Friston (2010) for a description of free-energy minimization as a general principle of brain function.

Equation (4.23) illustrates the traditional problem of Bayesian approaches – the integral in the denominator is often computationally intractable. As a model of human learning, Equation (4.23) would be more attractive if it included a plausible hypothesis about how humans could approximate such integrals sequentially in real time. As a start, it turns out that the posterior mean and precision can be rewritten (e.g., see Kruschke, 2011 for a derivation) as

$$\mu_{\mu_{A|w_n}} = \mu_{\mu_{A|w_{n-1}}} + \frac{\pi_A}{\pi_0 + \pi_A}(w_n - \mu_{\mu_{A|w_{n-1}}}) \tag{4.24}$$

and

$$\pi_{\mu_{A|w_n}} = \pi_0 + \pi_A. \tag{4.25}$$

Note that Equation (4.24) is in the same form as the iterative sample mean [Equation (4.5)], except that α is replaced with the ratio of the precision of the category A samples, π_A, to the sum of the category A precision plus the agent's precision about the task, π_0. Therefore, if category A precision is low (relative to π_0), then the learning rate is small. This makes sense intuitively – if we trust our model of the environment (i.e., π_0 is large), then we should be conservative about updating that model on the basis of noisy observations. On the other hand, if we have poor knowledge about the environment (i.e., π_0 is small) and there is not much variation in the samples (i.e., π_A is large), then we should use those samples to rapidly update our model of the environment.

This Bayesian formulation is beneficial for ensuring that prediction errors are precision-weighted according to their informativeness in a stable environment. However, this formulation will perform poorly in a nonstationary environment because the learning rate will not adapt to the environmental changes. For example, suppose the experimenter periodically changes the category A mean. The hierarchical Gaussian filter provides an efficient method for adapting to such changes in the environment by iteratively adjusting its estimate of μ_A using a variational Bayesian procedure (Mathys *et al.*, 2011, 2014).

The hierarchical Gaussian filter estimates μ_A in a hierarchical fashion. Let $x_1(n)$ denote the current estimate of μ_A on trial n [i.e., so on trial n, $\hat{\mu}_A = x_1(n)$]. Then the agent's model of category A on trial n is that

$$w_n \sim \mathcal{N}[x_1(n), \pi_A^{-1}], \tag{4.26}$$

since again, we are considering the simple case where π_A is known. This is the lowest level of the hierarchy. The next level up (i.e., level 2) estimates the distribution of the mean, $x_1(n)$. Specifically, the hierarchical Gaussian filter assumes that

$$x_1(n) \sim \mathcal{N}\{x_1(n-1),\ \exp[\kappa_1 x_2(n-1) + \omega_1]\}, \tag{4.27}$$

where κ_1 and ω_1 are constants, with $\kappa_1 > 0$, and $x_2(n)$ is a new random variable. The exponential function was chosen as a mathematically convenient form via which to estimate the variance of $x_1(n)$ (see, e.g., Mathys *et al.*, 2014). Because

$\kappa_1 > 0$, note that the variance of $x_1(n)$ increases with $x_2(n)$. The standard deviation of $x_1(n)$ is often referred to as volatility (Behrens *et al.*, 2007; Bland & Schaefer, 2012; Nassar *et al.*, 2010; Payzan-LeNestour & Bossaerts, 2011; R. C. Wilson, Nassar, & Gold, 2013), so $x_2(n)$ increases with volatility.

Level 3 of the hierarchy estimates the variance of $x_1(n)$ by assuming that

$$x_2(n) \sim \mathcal{N}\{x_2(n-1),\ \ \exp[\kappa_2 x_3(n-1) + \omega_2]\}, \tag{4.28}$$

where $x_3(n)$ is a new random variable that increases with the variance of volatility. In other words, $x_3(n)$ is measuring how much volatility is changing in the environment. In principle, this hierarchy can continue indefinitely. At each new level, the variance is defined in terms of a new random variable that is itself defined at the next higher level. So, for example, level 4 would define $x_3(n)$ as normally distributed with a variance that depends on a new random variable $x_4(n-1)$.

Another critical feature of the hierarchical Gaussian filter is that it specifies trial-by-trial update equations for the mean and precision parameters at each level of the hierarchy. These updates, which were all derived using a variational Bayesian procedure, are in the same general form as Equation (4.24), with the notable exception that the learning rates [i.e., the analog of $\pi_A/(\pi_0 + \pi_A)$ in Equation (4.24)] are sensitive to changes in the environment, including, for example, volatility and changes in volatility. For example, the update equations specify that when the environment becomes more volatile, the learning rate on μ_A increases. This makes sense intuitively because in a more volatile environment, deviations from our expectations may indicate that environmental events driving our sensory data have changed and learning should therefore proceed more rapidly.

The hierarchical Gaussian filter update equations enable real-time estimation of states and are optimal in the sense that they minimize variational free energy – an upper bound on an agent's surprise given its model of the world. The hierarchical Gaussian filter has a number of advantages over more traditional Bayesian models. First, it avoids the need to evaluate intractable integrals. Second, by placing different subject-specific priors on the κ_i and ω_i parameters, it provides a convenient method for modeling individual differences across agents. Third, it provides a foundation for RL-style update equations and firmly grounds RL models within the foundations of probability theory. Finally, the hierarchical Gaussian filter has also had considerable success at accounting for a wide variety of empirical phenomena, including impulsivity in healthy individuals (Paliwal *et al.*, 2014) and Parkinson's patients with deep brain implants (Paliwal *et al.*, 2018), reward-based decision-making in schizophrenia (Deserno *et al.*, 2020), social learning (Diaconescu *et al.*, 2017), perceptual learning (Weilnhammer *et al.*, 2018), and sensory learning (Iglesias *et al.*, 2013).

4.4.3 Supervised-Learning Models of Sensorimotor Adaptation

Models based on supervised learning are also popular. As described above, supervised learning is a local learning rule that uniquely changes each modifiable

weight or connection strength in the model. The most popular versions, which include backpropagation and the delta rule, implement a form of gradient descent (e.g., Rumelhart & McClelland, 1986). Consider a general model in which some unit i projects to some unit j. Let x_i denote the output of unit i, y_j denote the output of unit j, and denote the connection strength between units i and j by the parameter $\omega_{i,j}$. Then supervised learning algorithms change each $\omega_{i,j}$ differently. The most common approach, which is followed, for example, by gradient descent algorithms, is to modify $\omega_{i,j}$ according to the error between the desired output y_j^* of unit j and the observed output y_j. This error is typically referred to as $\delta_j = y_j^* - y_j$.

Gradient descent algorithms modify $\omega_{i,j}$ in a way that causes δ_j to decrease as quickly as possible at each time step. Specifically, if we let F represent the mathematical transformation that unit j performs on its input, then $y_j = F(x_i \mid \omega_{i,j})$. Gradient descent algorithms modify $\omega_{i,j}$ according to

$$\Delta\omega_{i,j} \propto -\frac{\partial \delta_j}{\partial \omega_{i,j}}, \tag{4.29}$$

that is, in proportion to the negative of the gradient on the error surface. Here, we can see that the key feature of a supervised learning system is that (1) the system is provided a teaching signal in the form of a desired output, and (2) the error signal (i.e., the difference between actual and desired output) is differentiable with respect to the parameters of the model. Equation (4.29) describes a local learning rule because every output unit in the model has its own unique desired output. Because of this, in response to an error signal at time t, some parameters will be increased, and others will be decreased. This property also strongly distinguishes supervised learning from RL, in which all active weights are either strengthened or weakened in accord with the presence or absence of unexpected rewards.

One prominent class of supervised-learning models uses linear dynamical systems to model the sensorimotor learning that causes adaptive changes in motor outputs in response to changing sensory inputs (Baddeley, Ingram, & Miall, 2003; Cheng & Sabes, 2006; Donchin, Francis, & Shadmehr, 2003; Scheidt, Dingwell, & Mussa-Ivaldi, 2001; Thoroughman & Shadmehr, 2000). Such changes are essential for coordinated and efficient execution of action selection and motor control. For example, as muscles are fatigued they require greater neural impulses to be activated, and therefore the motor commands that achieve some goal before muscle fatigue need to be scaled up to achieve that same goal after fatigue has accumulated. Sensorimotor learning also allows agents to adjust for noisy and dynamic environments. For example, the brakes on a rental car only feel foreign and jerky for a short while before we adapt our motor commands to smoothly operate them.

In the lab, sensorimotor learning is commonly studied with visuomotor adaptation experiments (Cunningham, 1989; Krakauer *et al.*, 2000; Martin *et al.*, 1996a, 1996b; Redding, Rossetti, & Wallace, 2005; Von Helmholtz, 1925). The agent's objective in such tasks is typically to reach from a start location to a target location as quickly, smoothly, and accurately as possible. After a baseline or familiarization

phase, the visual feedback provided by the moving hand is perturbed to introduce a mismatch between the actual and perceived hand position. Early experiments of this nature used prism glasses to induce lateral shifts, but recently the most common approach has been to use crude virtual-reality environments to impose visuomotor rotations such that movements beginning at the centre of a work space and traveling in a given direction generate on-screen cursor trajectories that match the radial distance from the reach origin but are rotated by some amount. People readily learn to compensate for a range of perturbations, quickly becoming proficient at moving to a target with relatively normal kinematics (Welch, 1986).

Since the early 2000s, linear dynamical systems endowed with supervised learning algorithms have provided a popular general model of sensorimotor learning, including behavior observed in visuomotor adaptation tasks (Baddeley, Ingram, & Miall, 2003; Cheng & Sabes, 2006; Donchin, Francis, & Shadmehr, 2003; Scheidt, Dingwell, & Mussa-Ivaldi, 2001; Thoroughman & Shadmehr, 2000). This is usually done by defining the state of the dynamical system as a sensorimotor transformation – that is, as an intermediate mapping from sensory input to motor output. Sensorimotor learning is then modeled as adaptive changes that reduce the errors in each state, and for this reason, models that employ this method are often referred to as state-space models.

As an example, consider a simple reaching task in which participants make center-out reaches to a single target. After some baseline phase in which participants are afforded the opportunity to familiarize themselves with the apparatus, the visual feedback is perturbed by a rotation. Further suppose that feedback is only given at the end of each reach, so that any adaptive change in the sensorimotor mapping occurs exclusively between trials. A simple and common state-space model of this task is described on trial n by the following equations:

$$\begin{aligned} \delta_n &= y_n^* - y_n, \\ x_{n+1} &= \beta x_n + \alpha \delta_n, \\ y_n &= x_n + \theta_n, \end{aligned} \tag{4.30}$$

where δ_n is the error (i.e., the angular distance between the reach endpoint and the target location), y_n^* is the desired output (e.g., the angular position of the reach target), y_n is the output and corresponds to the angle of the movement that will be generated when trying to reach to the target (i.e., it is a readout of the sensorimotor state), x_n is the state of the system (i.e., the sensorimotor transformation), β is a retention rate that describes how much is retained from the value of the state at the previous trial, α is a learning rate that describes how quickly states are updated in response to errors, and θ_n is the imposed rotation.

If we assume that in the absence of visuomotor rotations, the system is calibrated such that $\delta_n = 0$, and that this state corresponds to $x_n = 0$, then in the presence of a rotation $\theta_{n+1} \neq 0$, the system will experience the error $\delta_{n+1} = -\theta_{n+1}$, and will adjust x_{n+2} in a direction that would reduce the experienced error if the same rotation was applied on the next trial. For example, if θ_{n+1} is in a clockwise direction, then Equation (4.30) leads to $x_{n+2} = \beta x_{n+1} - \alpha \theta_{n+1}$. This means that

adaptation to a clockwise rotation occurs by adjusting the sensorimotor state to generate more counter-clockwise movements. If $\beta < 1$, then on each trial the state will respond to errors in the way just described, but will also return to baseline by some increment. Thus, in the absence of reach errors, the system has a tendency to reset itself. Because the goal of learning is to reduce the state error – that is, δ_n – this model is based on a form of supervised learning known as the delta rule or the Widrow–Hoff rule (Widrow & Hoff, 1960).

Another key feature of linear dynamical systems as models of sensorimotor learning is that they are easily modified to accommodate considerably more complexity than the simple version described above. For example, Cheng and Sabes (2006) outlined a more general form for these models governed by the following equations:

$$\begin{aligned} \mathbf{x}_n &= \mathrm{A}\mathbf{x}_{n-1} + \mathrm{B}\boldsymbol{\delta}_{n-1} + \boldsymbol{\eta}_{n-1}, \\ \mathbf{y}_n &= \mathrm{C}\mathbf{x}_n + \mathrm{D}\boldsymbol{\omega}_n + \boldsymbol{\gamma}_n, \end{aligned} \tag{4.31}$$

where $\mathbf{x}_n$ is a state vector of sensory transformations, $\boldsymbol{\delta}_n$ is the vector of errors – that is, the differences between the desired and actual states, $\boldsymbol{\eta}_n$ is a random vector that models noise in the learning process and is typically assumed to have a multivariate normal distribution with mean vector $\mathbf{0}$ and variance–covariance matrix Σ, $\mathbf{y}_n$ is a vector of motor outputs (e.g., angle and distance of movement), $\boldsymbol{\omega}_n$ is a vector of inputs to the system (e.g., θ_n in the simple example above), γ_n is a random vector that models noise in the output process (again typically assumed to have a multivariate normal distribution), and A, B, C, and D are all matrices of constant values. Note that this model modifies each state according to its own unique error, which is a hallmark of supervised learning (and of the delta rule).

In this form, it is easy to see that linear dynamical systems can be flexibly applied to a variety of sensorimotor learning scenarios in which the factors relevant to sensorimotor learning (stored in $\boldsymbol{\delta}_n$) can be stated independently of the factors relevant to sensorimotor output (stored in $\boldsymbol{\omega}_n$). The result is a convenient yet powerful framework that can be used to generate predictions about sensorimotor learning on a trial-by-trial basis, or even on a moment-to-moment basis if adaptive changes are thought to occur on that timescale. This approach is therefore well suited to modeling behavioral learning phenomena that change appreciably on these fast timescales.

4.5 Implementational Models of Human Learning

Implementational-level models explicitly state how neural circuits drive behavior, and how changes in connection weights within these circuits drive learning. Thus, at the core of these models are clear statements about the brain regions and networks that drive a behavior, and the forms of synaptic plasticity that govern changes in connection weights between neurons in constituent regions. We now know that synaptic plasticity comes in many different forms. For instance, it operates by different computational principles in different brain regions and

between different cell types (Doya, 2000; Feldman, 2009), and it is governed physiologically by different molecular mechanisms and intracellular signaling cascades. A complete review of both the physiological and computational underpinnings of every form of synaptic plasticity is well beyond the scope of this chapter. Instead, we focus on three forms of synaptic plasticity that are deeply understood from a physiological perspective, and are at the core of both classic and contemporary computational models of learning. In particular, we will discuss two-factor synaptic plasticity in the cerebral cortex and the hippocampus that is similar to Hebbian learning, three-factor DA-dependent synaptic plasticity in the basal ganglia that is similar to RL (Doya, 2000; Houk *et al.*, 1995), and a form of synaptic plasticity in the cerebellum that resembles supervised learning.

4.5.1 Physiology of DA-Dependent Two- and Three-Factor Synaptic Plasticity

The most common excitatory neurotransmitter in the brain is glutamate, and LTP at glutamatergic synapses is well understood. Glutamate binds to a number of different receptors, but the most important for LTP is NMDA. The biochemical details are not important for our purposes, except to note that NMDA requires partial depolarization to become activated, and so it has a higher threshold for activation than non-NMDA glutamate receptors. NMDA-receptor activation initiates a number of chemical cascades that can increase synaptic efficacy. Because of its high threshold, however, activation of NMDA receptors on the postsynaptic membrane requires strong presynaptic activation. If presynaptic activation either fails to activate or only weakly activates NMDA receptors, then a variety of evidence suggests that the long-term efficacy of the synapse is weakened (i.e., LTD occurs; Bear & Linden, 2001; Kemp & Bashir, 2001).

DA plays a critical modulatory role in these processes because it can potentiate synaptic efficacy if it is above baseline when NMDA receptors are activated, but synaptic weakening occurs if DA is below baseline during NMDA receptor activation (Calabresi *et al.*, 1996; Reynolds & Wickens, 2002; Yagishita *et al.*, 2014). A large literature shows that DA neurons in the ventral tegmental area and substantia nigra pars compacta increase their firing above baseline following unexpected rewards, and decrease their firing below baseline following the failure to receive an expected reward (e.g., Hollerman & Schultz, 1998; Mirenowicz & Schultz, 1994; Schultz, 1998). Thus, this form of DA-enhanced LTP should be in effect following an unexpected reward in any brain region that is a target of DA neurons. This includes the basal ganglia, the hippocampus, the amygdala, and all of the frontal cortex. In contrast, there is virtually no DA projection to visual or auditory cortex. In these regions, however, there is evidence that acetylcholine may play a modulatory role similar to DA in LTP and LTD (e.g., Gu, 2003; McCoy, Huang, & Philpot, 2009).

Although the biochemistry that mediates the modulatory role that DA plays in synaptic plasticity is similar in all DA target regions, the functional role of this

plasticity is qualitatively different in the striatum and the frontal cortex. Within the striatum, DA is quickly cleared from synapses by DA active transporter and, as a result, the temporal resolution of DA in the striatum is high enough for DA to serve as an effective trial-by-trial reinforcement-learning signal. For example, if the first response in a training session receives positive feedback and the second response receives negative feedback, then the elevated DA levels in the striatum that result from the positive feedback on trial 1 should have decayed back to baseline levels by the time of the response on trial 2. Unlike the striatum, however, the concentration of DA active transporter in the frontal cortex is low (e.g., Seamans & Robbins, 2010). As a result, cortical DA levels change slowly. For example, the delivery of a single food pellet to a hungry rat increases DA levels in the prefrontal cortex above baseline for approximately 30 min (Feenstra & Botterblom, 1996). Thus, the first rewarded behavior in a training session is likely to cause frontal cortical DA levels to rise, and the absence of DA active transporter will cause DA levels in the frontal cortex to remain high throughout the training session. As a result, all synapses that are activated during the session are likely to be strengthened, regardless of whether the associated behavior is appropriate or not. Thus, although DA may facilitate LTP in the frontal cortex, it appears to operate too slowly to serve as a frontal-cortical trial-by-trial reinforcement training signal (Lapish *et al.*, 2007).

From a computational perspective, the high temporal resolution of the striatal DA signal means that whether a synapse is strengthened or weakened depends on three factors: the amount of presynaptic activation, the amount of postsynaptic activation, and whether DA is above or below baseline. As a result, synaptic plasticity in the striatum is said to follow the *three-factor* learning rule (Wickens, 1993). In contrast, in the cortex, DA levels will change only slowly over time, so only two factors are needed to predict whether a synapse will be strengthened or weakened – the amount of pre- and postsynaptic activation. As a result, plasticity in the cortex follows the *two-factor* learning rule.

4.5.2 Models Based on Two-Factor Plasticity

Models of Two-Factor Plasticity

The structural changes at the synapse that accompany LTP and LTD are complex and highly diverse. For example, changes in synaptic plasticity might be mediated by changes in the number of receptors, their distribution, the type of receptors, or their sensitivity. But plasticity changes could also occur because of changes in the size and/or shape of dendritic spines. If our goal is to model learning-related changes in human behavior, then the molecular and cellular mechanisms that mediate changes in synaptic plasticity are irrelevant. We only need an accurate model of how *much* the efficacy of the synapse changes from one behavioral measurement to the next.

The structural changes at the synapse unfold continuously in time, but unless the behavioral measurements are continuous, there is no need to build a

continuous-time model. In particular, if the data have a discrete trial-by-trial structure, as is common in many cognitive-behavioral experiments, then a discrete-time model of changes in synaptic efficacy is often sufficient. Typically, such a model would be constructed from difference equations, where the index is trial number, so the implicit time interval is the duration of one trial. A continuous-time learning model (e.g., that uses differential equations) is typically required only when modeling a continuous-time behavioral task.

The simplest and original form of Hebbian learning predicts that between trials n and $n+1$, the strength of the synapse between units i and j, denoted by $w_{ij}(n+1)$, is

$$w_{ij}(n+1) = w_{ij}(n) + \alpha A_i(n) A_j(n), \tag{4.32}$$

where $A_i(n)$ and $A_j(n)$ are the total activations in units i and j on trial n and α is the learning rate. This model has two significant weaknesses. First, all terms in Equation (4.32) are positive, so this model includes no mechanism to weaken a synapse, and as a result, it cannot account for LTD. Second, note that it predicts that all synaptic strengths will eventually increase to infinity. For these reasons, a variety of alternative models of Hebbian learning have been proposed.

One model of two-factor plasticity, which can be seen as a generalization of classical Hebbian learning, assumes that (Ashby, 2018)

$$\begin{aligned} w_{ij}(n+1) = {} & w_{ij}(n) \\ & + \alpha\ \Delta\ H\left[A_j(n) - \theta_{\text{NMDA}}\right] A_i(n) \left\{1 - \mathrm{e}^{-\lambda\left[A_j(n) - \theta_{\text{NMDA}}\right]}\right\} \left[1 - w_{ij}(n)\right] \\ & - \beta\ H\left[\theta_{\text{NMDA}} - A_j(n)\right] A_i(n)\ \mathrm{e}^{-\lambda\left[\theta_{\text{NMDA}} - A_j(n)\right]} w_{ij}(n). \end{aligned} \tag{4.33}$$

The positive term describes conditions that strengthen the synapse and the negative term describes conditions that cause the synapse to be weakened. Ignore the constant Δ for now (i.e., assume $\Delta = 1$). The function $H[g(x)]$ is the Heaviside function that equals 1 when $g(x) > 0$ and 0 when $g(x) \leq 0$. The constant θ_{NMDA} represents the threshold for NMDA-receptor activation. Note that the synaptic strengthening term is positive only on trials when the postsynaptic activation exceeds the threshold for NMDA-receptor activation, and that the amount of strengthening depends on the product of the presynaptic activation and an exponentially increasing function of the postsynaptic activation. The $[1-w_{ij}(n)]$ term is a rate-limiting term that prevents $w_{ij}(n+1)$ from exceeding 1.0, and the constant λ scales the postsynaptic activation.

Note that the synapse is weakened only when the postsynaptic activation is below the NMDA threshold. Also note that the exponential term reaches its maximum when postsynaptic activation is near the NMDA threshold and decreases as the postsynaptic activation gets smaller and smaller. This is consistent with the neurobiology. For example, in the absence of any postsynaptic activation, we do not expect any synaptic plasticity. The $w_{ij}(n)$ at the end prevents $w_{ij}(n+1)$ from dropping below 0. Figure 4.1 shows predicted changes in synaptic strength

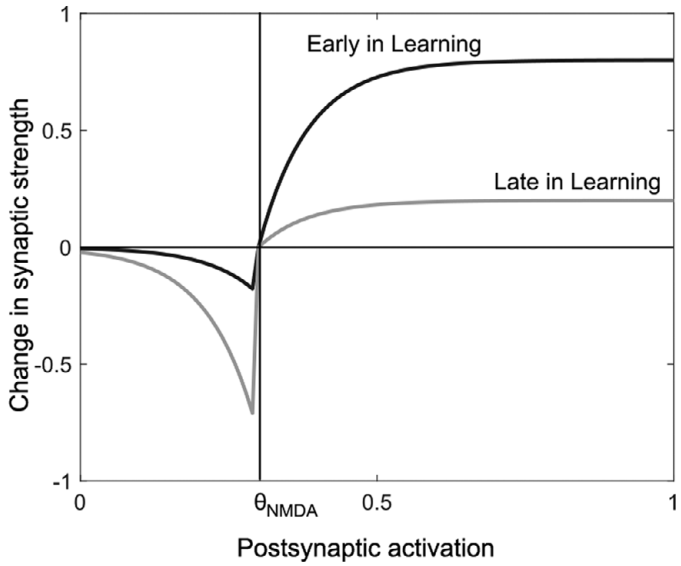


Figure 4.1 *Change in synaptic strength predicted by the two-factor learning model described in Equation (4.33) as a function of amount of postsynaptic activation (here scaled from 0 to 1). Predictions are shown for early in learning [i.e., when $w_{ij}(n) = 0.2$] and late in learning [i.e., when $w_{ij}(n) = 0.8$].*

[i.e., $w_{ij}(n + 1) - w_{ij}(n)$] for this model as a function of the magnitude of postsynaptic activation during both early [when $w_{ij}(n) = 0.2$] and late [when $w_{ij}(n) = 0.8$] learning.

The Equation (4.33) model of two-factor learning assumes that any activation in postsynaptic unit j was caused by activation in presynaptic unit i. This assumption is really only plausible in simple feedforward models. If unit j receives input from many other units, then Equation (4.33) could strengthen inappropriate synapses. In the mammalian brain, the magnitude and even the direction of plasticity at a synapse depends not only on the magnitude of the pre- and postsynaptic activations, but also on their timing – a phenomenon known as *spike-timing-dependent plasticity*. Considerable data show that if the postsynaptic neuron fires just after the presynaptic neuron then synaptic strengthening (i.e., LTP) occurs, whereas if the postsynaptic neuron fires first then the synapse is weakened (e.g., Bi & Poo, 2001; Sjöström *et al.*, 2008). Furthermore, the magnitude of both effects seems to fall off exponentially as the delay between the spikes in the pre- and postsynaptic neurons increases. Let T_{pre} and T_{post} denote the time at which the pre- and postsynaptic units fire, respectively. Then a popular model of spike-timing-dependent plasticity (e.g., Zhang *et al.*, 1998) assumes that the amount of change in the synaptic strength is

$$\Delta = \begin{cases} e^{-\theta_+(T_{\text{post}} - T_{\text{pre}})}, & \text{if } T_{\text{post}} > T_{\text{pre}} \\ e^{\theta_-(T_{\text{post}} - T_{\text{pre}})}, & \text{if } T_{\text{post}} < T_{\text{pre}} \end{cases}, \tag{4.34}$$

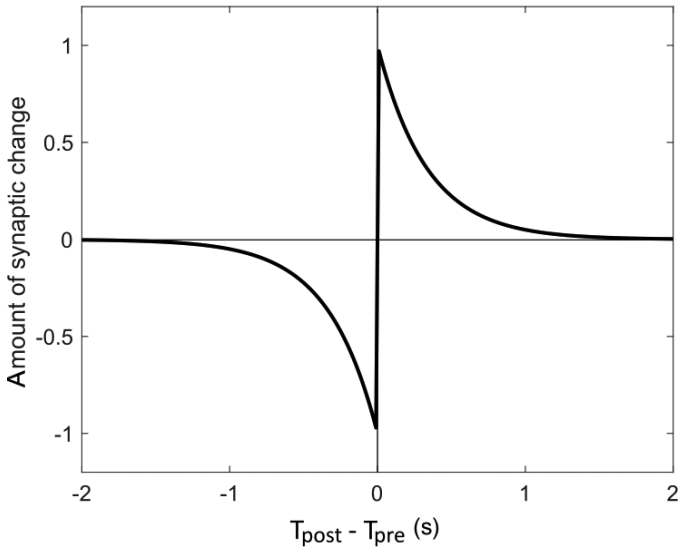


Figure 4.2 *Amount of change in synaptic strength predicted by spike-timing-dependent plasticity as a function of the difference in time between firing in the postsynaptic neuron (i.e., T_{post}) and the presynaptic neuron (i.e., T_{pre}).*

where θ_+ and θ_- are parameters that determine the decay rates of synaptic strengthening and weakening, respectively. Figure 4.2 shows an example of this function.

To incorporate spike-timing-dependent plasticity into two-factor learning, the first step is to compute Δ from Equation (4.34) anytime the pre- and postsynaptic units both fire. Next this value is inserted into Equation (4.33) to compute $w(n+1)$.

Models of Human Learning that Incorporate Two-Factor Plasticity

Hasselmo and Wyble (1997) proposed a model that includes two-factor plasticity in the hippocampus to account for the effects of scopolamine, an acetylcholine anatagonist, on free recall and recognition. They tested this model against data from an experiment reported by Ghoneim and Mewaldt (1975), in which participants studied lists of 16 words each and were then tested on their ability to recall and recognize the studied words. Recall and recognition were both intact when scopolamine was administered between study and test. In contrast, the administration of scopolamine before study impaired recall, but not recognition.

Figure 4.3 shows the neural architecture of the Hasselmo and Wyble (1997) model. Neural activation in each region was modeled by firing-rate models (e.g., see Ashby, 2018). The hippocampus contains two subfields, the cornus ammonis and the dentate gyrus, each of which receives input from the entorhinal cortex, which in turn is driven by widespread input from the neocortex. The network is

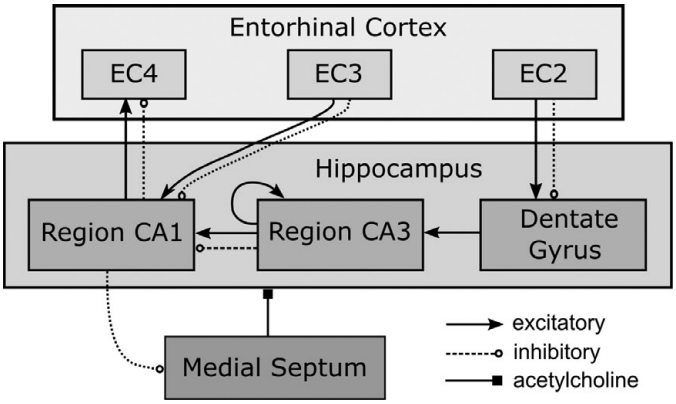


Figure 4.3 *The neural architecture of the Hasselmo and Wyble (1997) hippocampal model. EC2, EC3, and EC4 denote different subregions in the entorhinal cortex, whereas CA1 and CA3 denote different subregions in the cornus ammonis. Two-factor learning occurs at virtually all synapses, except at the synapses between dentate gyrus and CA3 and between CA1 and medial septum.*

characterized by sparse encoding and many feedback loops, and the behavior of the model is governed largely by how the resulting network dynamics approach attractor states.

The network has two global states (encoding and retrieval) that are controlled by the concentration of acetylcholine. The encoding mode is triggered by elevated acetylcholine and is characterized by potentiated two-factor learning at all plastic synapses (hence encoding), and also by inhibited output from EC4 back to neocortex (hence no retrieval). Acetylcholine can also reduce excitatory transmission, limiting the effects of recurrent collaterals and making the network primarily sensitive to external inputs. This is good for learning because it helps reduce interference between new items and previously stored items. The retrieval mode is triggered by depressed acetylcholine and is characterized by reduced two-factor learning at plastic synapses (hence no encoding) and also by potentiated output from EC4 to neocortex (hence retrieval). Low acetylcholine also allows excitatory transmission via the network's recurrent collaterals, making the network sensitive to stored representations.

The form of two-factor learning used in the model is essentially the same as in Equation (4.33), but with the addition of providing a model of how the α and β parameters change with concentrations of acetylcholine. The model successfully simulates recall when context (i.e., cues associated with the word list) is presented to the network and it outputs words associated with that context. Additionally, the model successfully simulates recognition when it is presented with words and it outputs the context associated with the words. Hasselmo and Wyble (1997) showed that in the presence of scopolamine, the network has no difficulty retrieving inputs learned prior to scopolamine administration, whereas recall of inputs encoded in

the presence of scopolamine is disrupted and recognition of these inputs is spared. For a full explanation of the network dynamics that enable the model to account for these phenomena, see Hasselmo and Wyble (1997).

Here we only focus on the synaptic effects of acetylcholine on the hippocampus. However, Hasselmo and Wyble (1997) also explored the effects on depolarization and adaptation of neurons. Furthermore, the model was also shown to account for the list length and list strength effects (Murdock & Kahana, 1993; Murdock Jr., 1962; Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Roberts, 1972) in addition to making predictions about the effects of scopolamine on paired-associate tasks (Caine *et al.*, 1981; Crow & Grove-White, 1973; Ostfeld & Aruguete, 1962). The Hasselmo and Wyble (1997) model provides a good illustration of how relatively simple two-factor plasticity rules can be incorporated into sophisticated implementational-level models that account for neuropharmacological and behavioral phenomena.

4.5.3 Models Based on DA-Dependent Three-Factor Plasticity

Models of DA-Dependent Three-Factor Plasticity

In the striatum, DA reuptake is fast, so plasticity follows the three-factor rule. In other words, three factors are needed to strengthen a synapse: strong presynaptic activation, strong postsynaptic activation, and DA above baseline. If any of these factors are missing, then the synapse is weakened. A discrete-time model of three-factor learning is as follows:

$$\begin{aligned} w_{ij}(n+1) = w_{ij}(n) & \\ & + \alpha\, H\left[A_j(n) - \theta_{\text{NMDA}}\right] H[D(n) - D_{\text{base}}] \\ & \times A_i(n) \left\{1 - \mathrm{e}^{-\lambda[A_j(n) - \theta_{\text{NMDA}}]}\right\} [D(n) - D_{\text{base}}][1 - w_{ij}(n)] \\ & - \beta\, H\left[A_j(n) - \theta_{\text{NMDA}}\right] H[D_{\text{base}} - D(n)] \\ & \times A_i(t) \left\{1 - \mathrm{e}^{-\lambda[A_j(n) - \theta_{\text{NMDA}}]}\right\} [D_{\text{base}} - D(n)] w_{ij}(n) \\ & - \gamma\, H\left[\theta_{\text{NMDA}} - A_j(n)\right] A_i(n)\, \mathrm{e}^{-[\theta_{\text{NMDA}} - A_j(n)]} w_{ij}(n), \end{aligned} \tag{4.35}$$

where $D(n)$ is the amount of DA released on trial n and D_{base} is the baseline DA level (Ashby, 2018).

Recall that $H(x)$ is the Heaviside function, which equals 0 if $x \leq 0$ and 1 if $x > 0$. Therefore, the positive LTP term equals 0 except when presynaptic activation exceeds the postsynaptic NMDA threshold [i.e., $A_j(n) > \theta_{\text{NMDA}}$] and DA exceeds baseline [i.e., $D(n) > D_{\text{base}}$]. Thus, synaptic strengthening requires three conditions – strong presynaptic activation, postsynaptic activation above the threshold for NMDA-receptor activation, and DA above baseline. Once these conditions are met, synaptic strengthening is the same as in the Equation (4.33) two-factor learning model. Two different conditions cause the synapse to be weakened. The second [the last γ term in Equation (4.35)] is the same as in the two-factor model.

The first (i.e., the β term), however, is unique to striatal-mediated three-factor plasticity. Cortical–striatal synapses are weakened if postsynaptic activation is strong and DA is below baseline – a condition that would occur, for example, on trials when feedback indicates the trial n response was incorrect.

The Equation (4.35) model of three-factor plasticity requires that we specify the amount of DA released on every trial in response to the feedback signal [the $D(n)$ term]. The more that DA increases above baseline (D_{base}), the greater the increase in synaptic strength, and the more it falls below baseline, the greater the decrease.

Although there are a number of powerful models of DA release, Equation (4.35) requires only that we specify the amount of DA released to the feedback signal on each trial. The key empirical results are (e.g., Schultz, Dayan, & Montague, 1997; Tobler, Dickinson, & Schultz, 2003): (1) midbrain DA neurons fire tonically, and therefore have a nonzero baseline (i.e., spontaneous firing rate); (2) DA release increases above baseline following unexpected reward, and the more unexpected the reward the greater the release, and (3) DA release decreases below baseline following unexpected absence of reward, and the more unexpected the absence, the greater the decrease. One common interpretation of these results is that over a wide range, DA firing is proportional to the reward prediction error (RPE) – that is, to the difference between obtained reward and predicted reward. If we denote the obtained reward on trial n by r_n and the predicted reward by P_n, then the RPE on trial n is defined as

$$RPE_n = r_n - P_n. \tag{4.36}$$

So positive prediction errors occur when the reward is better than expected, and negative prediction errors when the reward is worse than expected. Note that either a positive or negative prediction error is a signal that learning is incomplete.

A simple model of DA release can be built by specifying how to compute (1) obtained reward, (2) predicted reward, and (3) exactly how the amount of DA release is related to the RPE. A straightforward solution to these three problems is as follows (Ashby & Crossley, 2011). First, in tasks that provide positive feedback, negative feedback, or no feedback on every trial and where reward magnitude never varies, a simple model can be used to compute obtained reward. Specifically, define the obtained reward r_n on trial n as +1 if correct or reward feedback is received, 0 in the absence of feedback, and -1 if error feedback is received.

Second, following an old tradition (Bush & Mosteller, 1951), predicted reward can be computed using the iterative sample mean [i.e., Equation (4.5)]:

$$P_{n+1} = P_n + \alpha_p(r_n - P_n), \tag{4.37}$$

where α_p is the learning rate.[4]

4 The subscript p is to distinguish this learning rate parameter from the learning rate α in Equation (4.35).

The final step is to compute the amount of DA released for any specific value of RPE. A simple model, which is consistent with the single-unit recording data reported by Bayer and Glimcher (2005), assumes that

$$D(n) = \begin{cases} 1, & \text{if } RPE > 1 \\ 0.8\, RPE + 0.2, & \text{if } -0.25 < RPE \leq 1. \\ 0 & \text{if } RPE < 0.25 \end{cases} \tag{4.38}$$

Note that this model assumes a baseline DA level of 0.2 [i.e., $D(n)$ on trials when $RPE = 0$]. Positive RPEs increase DA release above this baseline, and negative RPEs depress it below baseline.

Figure 4.4 shows predicted changes in synaptic strength [i.e., $w_{ij}(n+1) - w_{ij}(n)$] for this model as a function of the magnitude of postsynaptic activation, separately for early [when $w_{ij}(n) = 0.2$] and late learning [when $w_{ij}(n) = 0.8$], and following correct and incorrect responses. Note that synaptic plasticity following correct (rewarded) responses is similar to plasticity in the two-factor model (compare the top panel of Figure 4.4 with Figure 4.1). The only real difference is that plasticity is attenuated more during late learning in the three-factor model. This is because DA fluctuations decrease as rewards become more predictable. Note also that errors have a greater effect on synaptic plasticity late in learning. This is because errors are expected early in learning, so DA fluctuations are small. Late in learning, however, when accuracy is high, errors are unexpected, which causes a large DA depression and therefore a large decrease in synaptic efficacy.

Relationship of Three-Factor Plasticity to Psychological Constructs of RL

Three-factor plasticity may – in some respects – be seen as a possible neural implementation of the many SR association learning models that were inspired by Thorndike's (1927) law of effect. The obvious analogy maps presynaptic activity onto the stimulus component, postsynaptic activity onto the response component, and DA onto the reinforcement signal. A step further, and we might expect the stimulus component to be encoded by a primary sensory neuron, the response unit to be encoded by a primary motor neuron, and the reinforcement signal to strengthen or weaken the synapse between these two neurons. Although human neuroanatomy supports the existence of direct projections from sensory to motor areas, the evidence suggests that these synapses are not strengthened via a DA-mediated reinforcement signal, because DA reuptake in the cortex is too slow. Rather, the available evidence suggests that sensory and motor neurons are indirectly wired together via a DA-mediated reinforcement signal in the basal ganglia. Here, stimulus–response associations can be learned at cortical–striatal synapses, with the striatum projecting via a multisynaptic pathway to the motor neurons representing the response component of the association. From this perspective, the anatomy and physiology of cortical–basal ganglia–DA interactions may provide a plausible neural substrate for the classic psychological constructs of stimulus–response learning originally posed by Thorndike. However, the anatomy

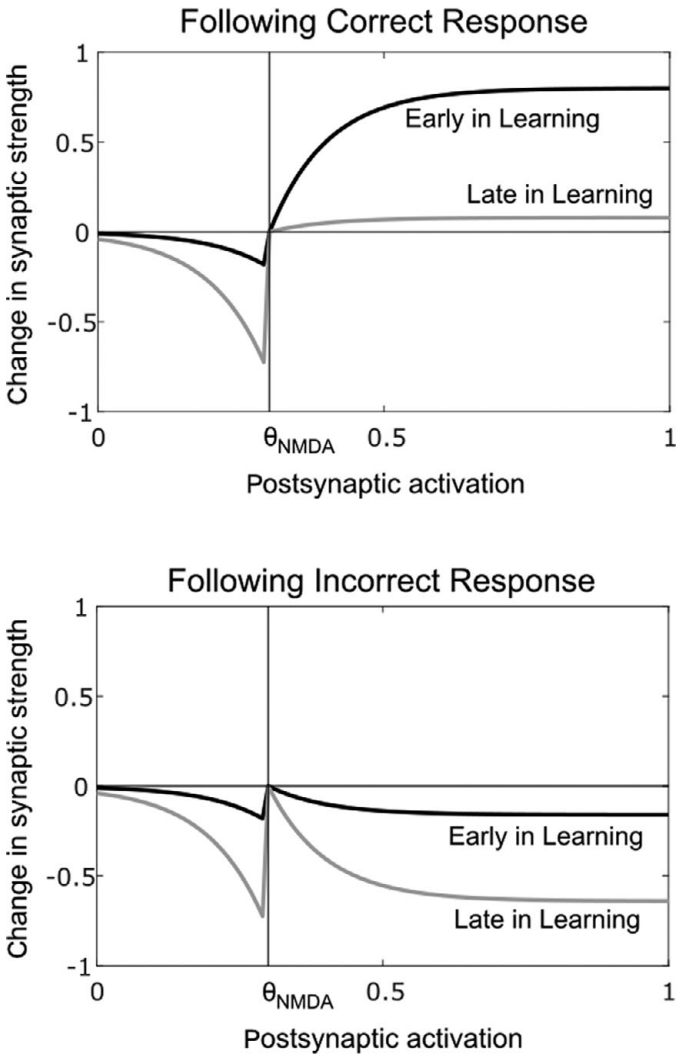


Figure 4.4 *Changes in synaptic strength predicted by the model of three-factor plasticity described in Eq. 4.35 as a function of amount of postsynaptic activation (here scaled from 0 to 1). Predictions are shown for early in learning [i.e., when* $w_{A,B}(n) = 0.2$*] and late in learning [i.e., when* $w_{A,B}(n) = 0.8$*], and following feedback that the response was correct or incorrect. (*$\alpha = 2$, $\beta = 4$, $\gamma = 1$.*)*

also suggests that the association mechanism is more indirect and complex than in the original proposals of direct reinforcement of stimulus–response components.

Relationship of Three-Factor Plasticity to Machine-Learning Constructs of RL

Three-factor plasticity in the basal ganglia may also offer a plausible biological substrate for various machine-learning constructs of RL. In this view,

cortical–striatal synaptic weights implement a value function, and DA neurons provide the reinforcement signal – a role motivated by the finding that DA neuron firing reflects an RPE (Glimcher, 2011; Schultz, Dayan, & Montague, 1997). This arrangement could be seen as compatible with a range of specific RL algorithms, including temporal-difference learning, Q learning, and actor–critic architectures, although the mapping does not seem perfect for any of these.

To be compatible with temporal-difference learning, cortical–striatal synaptic weights would need to encode a value function that depends exclusively on sensory states (i.e., is independent of action). This sort of value function encoding may be characteristic of the ventral striatum (e.g., nucleus accumbens). The value function would also need to be used to generate prediction errors, which is consistent with one of the roles sometimes ascribed to the ventral striatum. However, the value function would also need to operate under the assumption of a fixed action policy and, at present, it is unclear whether the ventral striatum learns different value functions for different policies. Another feature of temporal-difference learning, which makes it a problematic model of DA neuron firing, is that, as we saw earlier, the temporal-difference signal propagates back one time step every trial, until it reaches the cue, at which point the propagation ends. DA neurons initially fire to the reward, and eventually, after learning occurs, they begin to fire to the cue. But there is no evidence that the propagation backwards is incremental – that is, there is never a DA response to an intermediate time point between cue and reward.[5]

To be compatible with Q learning, cortical–striatal synaptic weights would need to encode a value function that combines both sensory states and actions. This sort of value function encoding may be characteristic of the dorsal striatum. Parts of the dorsal striatum have quite direct access to motor areas of cortex, so it is plausible that they could also directly implement the action-selection components of Q learning. However, DA-encoded RPEs would also need to be derived from the value estimates provided by the dorsal striatum. At present, it is unclear to what degree such prediction errors factor in information about action.

In actor–critic RL models, an actor system implements an action selection policy, and a critic system estimates the value of different states and uses these estimates to generate prediction errors, which are then used to update the critic's value estimates and the actor's selection policy. Of all the machine-learning RL algorithms, these models may most easily map onto three-factor plasticity in the basal ganglia (Houk, Adams, & Barto, 1995; Joel *et al.*, 2002; Sutton & Barto, 1998). In this view, the critic is implemented by the DA system and the actor is implemented by cortical–striatal projections through the dorsal striatum. Since the critic is a separate module from the actor, there is no need for cortical–striatal synaptic weights (part of the actor) to be used to compute prediction errors. However, this view does not say where and how the value function is implemented. One possibility is the ventral striatum (Takahashi, Schoenbaum, & Niv, 2008).

5 This problem can be solved by replacing the temporal-difference learning algorithm with a version that includes an eligibility trace, which allows the error to propagate backwards by more than a single state per step (Sutton & Barto, 1998).

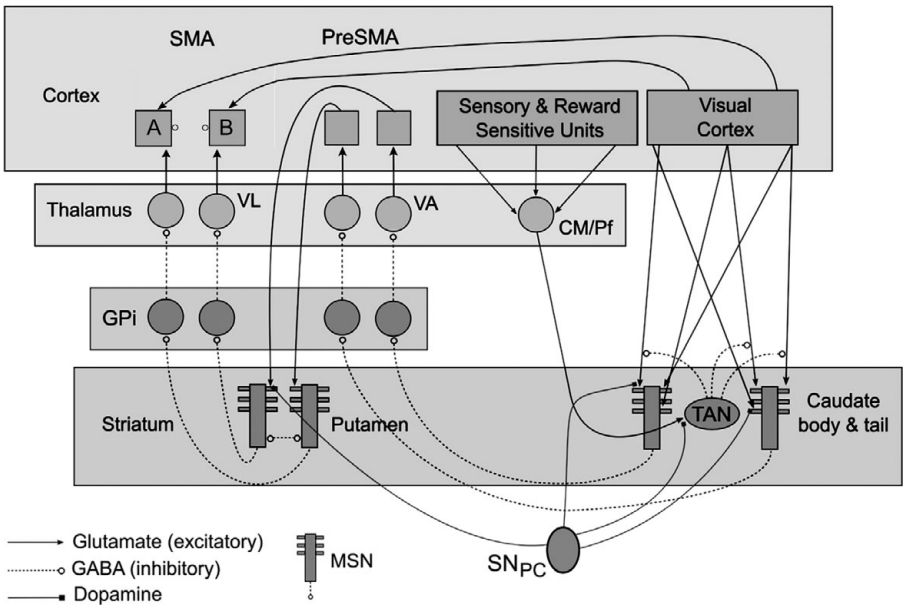


Figure 4.5 *The neural architecture of the COVIS model of procedural learning for a two-alternative forced-choice task with responses A and B (SMA = supplementary motor area, PreSMA = presupplementary motor area, VL = ventral lateral nucleus of the thalamus, VA = ventral anterior nucleus of the thalamus, CM/Pf = centromedian and parafascicular nuclei of the thalamus, GPi = internal segment of the globus pallidus, TAN = tonically active neuron, SN_{PC} = substantia nigra pars compacta, MSN = medium spiny neuron of the striatum).*

Models of Human Learning that Incorporate Three-Factor Plasticity

The COVIS procedural-learning model incrementally learns arbitrary stimulus–response associations via a model of three-factor plasticity that is essentially identical to Equation (4.35). Figure 4.5 shows the architecture of the model (Ashby *et al.*, 1998; Ashby & Crossley, 2011; Ashby & Waldron, 1999; Cantwell, Crossley, & Ashby, 2015). The key structure is the striatum, a major input region within the basal ganglia that includes the caudate nucleus and the putamen. In primates, all of the extrastriate visual cortex projects directly to the striatum, with a cortical–striatal convergence ratio of approximately 10,000 to 1 (e.g., C. J. Wilson, 1995). The model assumes that, through a procedural-learning process, each striatal medium spiny neuron associates a motor goal (e.g., press the button on the left) with a large group of visual cortical neurons (i.e., all that project to it). Much evidence supports the hypothesis that procedural learning is mediated within the basal ganglia, and especially at cortical–striatal synapses, which exhibit three-factor plasticity (Ashby & Ennis, 2006; Houk, Adams, & Barto, 1995; Mishkin, Malamut, & Bachevalier, 1984; Willingham, 1998). The COVIS procedural-learning model is a formal instantiation of these ideas.

Note that the model includes two loops through the basal ganglia (Cantwell, Crossley, & Ashby, 2015). One loop projects from the visual cortex through the body and tail of the caudate nucleus and terminates in the presupplementary motor area, and the second loop projects from the presupplementary motor area through the putamen and terminates in the supplementary motor area. Because this second loop terminates in the premotor cortex, COVIS predicts that the associations that are learned are between stimuli and motor goals. Both loops rely on three-factor learning at cortical–striatal synapses. The first loop learns which stimuli are associated with the same response and the second loop learns what motor response is associated with each of these stimulus clusters. In a novel task, both types of learning are required. However, note that if we train agents to make accurate classification responses and then switch the responses associated with the two stimulus classes, then the classes remain unchanged – only the response mappings must be relearned. So COVIS predicts that reversing the locations of the response keys will interfere with procedural classification performance, but that recovery from such a reversal should be easier than novel classification learning – a prediction that has been supported in several studies (Cantwell, Crossley, & Ashby, 2015; Kruschke, 1996; Maddox *et al.*, 2010; Sanders, 1971; Wills *et al.*, 2006).

COVIS uses a biologically accurate model of spiking in individual neurons proposed by Izhikevich (2003). Let $V_i(t)$ and $V_j(t)$ denote the intracellular voltages of the pre- and postsynaptic neurons, respectively, at time t. Then the Izhikevich (2003) model assumes that the intracellular voltage of the postsynaptic neuron on trial n is described by the following differential equations:

$$\begin{aligned}\frac{dV_j(t)}{dt} &= w_{ij}(n)f\left[V_i(t)\right] + \beta + \gamma\left[V_j(t) - V_r\right]\left[V_j(t) - V_t\right] - \theta U_j(t),\\ \frac{dU_j(t)}{dt} &= \lambda\left[V_j(t) - V_r\right] - \omega U_j(t),\end{aligned} \tag{4.39}$$

where β, γ, V_r, V_t, θ, λ, and ω are constants that are adjusted to produce dynamical behavior that matches the neural population being modeled. $U_j(t)$ is an abstract regulatory term that is meant to describe slow recovery in the postsynaptic neuron after an action potential is generated. Equation (4.39) produces the upstroke of an action potential via its own dynamics. To produce the downstroke, $V_j(t)$ is reset to V_{reset} when it reaches V_{peak}, and at the same time, $U_j(t)$ is reset to $U_j(t) + U_{\text{reset}}$, where V_{reset}, V_{peak}, and U_{reset} are free parameters.

The model has many free parameters and therefore can fit a wide variety of dynamical behavior. Izhikevich (2003) identified different sets of parameter values that allow the model to mimic the spiking behavior of approximately 20 different types of neurons. For example, one set of parameter values allows the model to mimic the firing properties of the striatal medium spiny neurons shown in Figure 4.5 (including, e.g., their up and down states), and another set of values allows the model to mimic the regular spiking neurons that are common in the cortex. Furthermore, Ashby and Crossley (2011) modified the Izhikevich model to account for the unusual dynamics of the striatal cholinergic interneurons (which

produce a pronounced pause in their high-tonic firing rate following excitatory input). In all these cases, the parameters are fixed by fitting the model to single-unit recording data from the neural population being modeled. Once set, the parameter values that define the models of each individual neuron type then remain fixed throughout all applications. Therefore, when testing the model against behavioral or neuroimaging data, the models of each neuron type have zero free parameters.

The function $f[V_i(t)]$ in Equation (4.39) models the input from the presynaptic neuron i. In particular, it uses a simple model called the alpha function to mimic the temporal delays of spike propagation and the temporal smearing that occurs at the synapse (Rall, 1967). Specifically, the alpha function assumes that every time the presynaptic neuron spikes, the following input is delivered to the postsynaptic neuron (with spiking time $t = 0$):

$$\alpha(t) = \frac{t}{\delta} \exp\left(\frac{\delta - t}{\delta}\right), \tag{4.40}$$

where δ is a constant. This function has a maximum value of 1.0 and it decays to 0.01 at $t = 7.64\delta$. Thus, δ can be chosen to model any desired temporal delay. Suppose the presynaptic neuron i produces N spikes that occur at times t_1, $t_2, \ldots, t_N$. Then the function f in Equation (4.39) is

$$f[V_i(t)] = \sum_{k=1}^{N} [\alpha(t - t_k)]^+, \tag{4.41}$$

where

$$[\alpha(t - t_k)]^+ = \begin{cases} \alpha(t - t_k) & \text{if } t > t_k \\ 0 & \text{if } t \leq t_k \end{cases}. \tag{4.42}$$

Finally, synaptic plasticity, and therefore learning, is modeled by the $w_{ij}(n)$ multiplier on $f[V_i(t)]$ in Equation (4.39). The value of this term is adjusted trial-by-trial, either via the two-factor [Equation (4.33)] or three-factor [Equation (4.35)] models of synaptic plasticity. COVIS assumes that the procedural learning in the striatum is mediated by three-factor plasticity at cortical–striatal synapses. Therefore, the presynaptic neuron i in Equation (4.39) would be in the cortex (either visual cortex or presupplementary motor area), the postsynaptic neuron j would be a medium spiny neuron in the striatum, and $w_{ij}(n)$ would be adjusted trial-by-trial by Equation (4.35). For a complete description of this type of mathematical modeling, called computational cognitive neuroscience, see Ashby (2018).

COVIS uses the Izhikevich (2003) model [i.e., Equation (4.39)] to model spiking in all neuron types shown in all brain regions illustrated in Figure 4.5, and it uses the alpha function [Equation (4.41)] to model synaptic transmission between all connected neurons. The supplementary motor area in the model includes as many simulated neurons as there are response alternatives in the task under study. Figure 4.5 shows the architecture of the model when applied to a two-alternative

forced-choice task with responses A and B. To generate a motor behavior, a response threshold is set on the integrated alpha function of each supplementary motor area unit [i.e., the integral of Equation (4.41)]. The first unit to exceed its threshold initiates its associated motor response. The lateral inhibition between competing supplementary motor area units causes the units to display the type of push–pull activity identified in many premotor regions of the cortex (e.g., as in Shadlen & Newsome, 2001). Formally, this architecture – that is, separate accumulators with lateral inhibition – mimics a drift diffusion process, but of course it is more easily extended to tasks with more than two response alternatives (Bogacz *et al.*, 2007; P. L. Smith & Ratcliff, 2004; Usher & McClelland, 2001).

Note that COVIS predicts that synaptic strengthening can only occur when the visual trace of the stimulus and the postsynaptic effects of DA overlap in time. More specifically, synaptic plasticity in the striatum is strongest when the intracellular signaling cascades driven by NMDA receptor activation and DA D1 receptor activation coincide (Lisman, Schulman, & Cline, 2002). The further apart in time these two cascades peak, the less effect DA will have on synaptic plasticity. For example, Yagishita *et al.* (2014) reported that synaptic plasticity was best (i.e., greatest increase in spine volume on striatal medium spiny neurons) when DA neurons were stimulated 600 ms after medium spiny neurons. When the DA neurons were stimulated before or 5 s after the medium spiny neurons, then no evidence of any plasticity was observed. In a task mediated by procedural learning, activation of the medium spiny neurons should occur just before the motor response, and activation of the DA neurons should occur just after the feedback. So COVIS predicts that feedback delays during procedural learning should have effects that are similar to those observed by Yagishita *et al.* (2014). In fact, many studies have confirmed this prediction in a form of category learning thought to depend on procedural learning (i.e., the information-integration categorization task; Dunn, Newell, & Kalish 2012; Maddox, Ashby, & Bohil 2003; Maddox & Ing, 2005; Worthy, Markman, & Maddox, 2013). Valentin, Maddox, and Ashby (2014) showed that the COVIS procedural-learning model can accurately account for the effects of all these feedback delays. In contrast, the same studies showed that delays up to 10 s have no effect on rule-based category learning that is thought to be mediated primarily in the prefrontal cortex.

Ashby and Crossley (2011) proposed that the striatal cholinergic interneurons serve as a context-sensitive gate between cortex and striatum (see also Crossley, Ashby, & Maddox, 2013, 2014; Crossley *et al.*, 2016). The idea, which is supported by a wide variety of neuroscience evidence, is that the striatal cholinergic interneurons tonically inhibit cortical input to striatal medium spiny neurons (e.g., Apicella, Legallet, & Trouche, 1997; Pakhotin & Bracci, 2007). The striatal cholinergic interneurons are driven by neurons in the centremedian–parafascicular nuclei of the thalamus, which in turn are broadly tuned to features of the environment. In rewarding environments, the cholinergic interneurons learn to pause to stimuli that predict reward, which releases the cortical input to the striatum from inhibition. This allows striatal output neurons to respond to excitatory cortical input,

thereby facilitating cortical–striatal plasticity. In this way, cholinergic interneuron pauses facilitate the learning and expression of striatal-dependent behaviors. When rewards are no longer available, the cholinergic interneurons cease to pause, which prevents striatal-dependent responding and protects striatal learning from decay.

Extending the COVIS procedural-learning system to include striatal cholinergic interneurons allows the model to account for many new phenomena – some of which have posed difficult challenges for previous learning theories. One of these is that the reacquisition of an instrumental behavior after it has been extinguished is considerably faster than during original acquisition (Ashby & Crossley, 2011). The model accounts for this ubiquitous phenomenon because the withholding of rewards during the extinction period causes the cholinergic interneurons to stop pausing to sensory cues in the conditioning environment (since they are no longer associated with reward). This closes the gate between the cortex and the striatum, which prevents further weakening of the cortical–striatal synapses. When the rewards are reintroduced, the cholinergic interneurons relearn to pause, and the behavior immediately reappears because of the preserved synaptic strengths.

4.5.4 Models Based on Plasticity that Mimics Supervised Learning

The cerebellum is commonly thought to provide a neural substrate for supervised learning (Doya, 1999) and there is a rich basis of implementational-level models in support of this view, beginning with the seminal work of Marr (1969). For this reason, the following sections are focused on learning in the cerebellum.

Learning in the Cerebellum

The cerebellum is anatomically arranged into multisynaptic loops with the cerebral cortex (Ramnani, 2006). Influence over the cerebellum is orchestrated through the pons, which receives widespread inputs from cortical and peripheral sites – including those associated with proprioception (Sawtell, 2010), haptics (Ebner & Pasalar, 2008; Shadmehr & Krakauer, 2008; Weiss & Flanders, 2011), and ongoing motor commands (Schweighofer *et al.*, 1998) – and gives rise to the mossy fiber inputs to cerebellar granule cells. Granule cells give rise to parallel fibers, which provide one of two major inputs to the Purkinje cells of the cerebellar cortex, which are the only projection neurons in the cerebellar cortex. The second input to the Purkinje cells comes from climbing fibers, which originate in the inferior olive. Purkinje cells project to the cerebellar deep nuclei, which in turn are relayed to the thalamus, and ultimately back to the cortex, thereby closing the anatomical loop.

Classic theories proposed that the cerebellum uses a form of supervised learning to control and coordinate motor function (Albus, 1971; Ito, 1984; Marr, 1969). In essence, these theories viewed the cerebellum as a biological implementation of a perceptron (Rosenblatt, 1958; see Figure 4.6), with distributed inputs provided

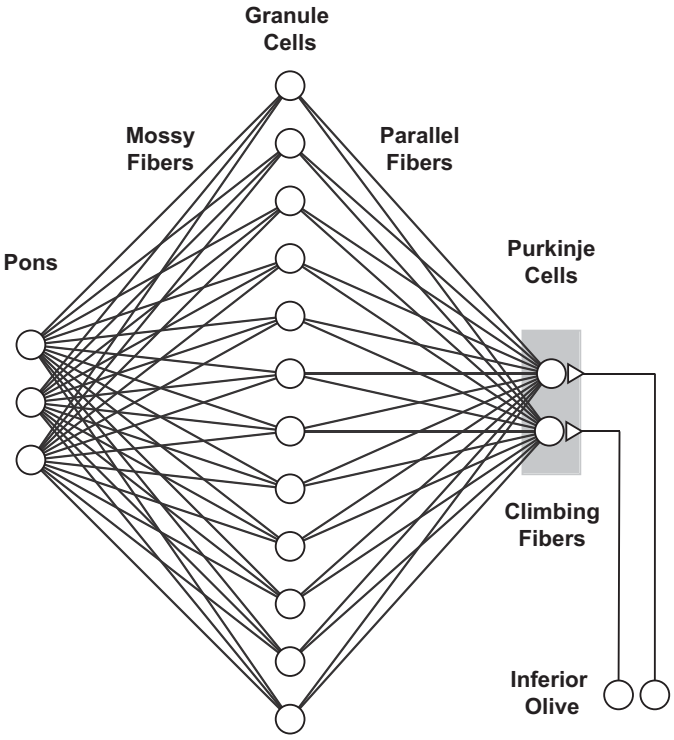


Figure 4.6 *Simplified neuroanatomy of the cerebellum when viewed as a three-layer perceptron. Purkinje cell output is inhibitory. All other illustrated projections are excitatory. See text for further details.*

by the mossy fibers, error signals communicated by the climbing fibers, and supervised learning carried out by synaptic plasticity at the synapses between parallel fibers and Purkinje cells (either LTP as originally proposed by Marr, 1969 or LTD as originally proposed by Ito, 1984). Ito and colleagues played pivotal roles in establishing the biological plausibility of this synaptic plasticity (e.g., Ito, 1984).

The anatomy of the cerebellum is unique in a few ways that probably played a large role in the development of these models. First, granule cells constitute more than half the neurons in the mammalian cerebellum (Eccles, Ito, & Szentágothai, 1967; Palay & Chan-Palay, 2012), so mossy fiber input seems like a plausible biological substrate for the distributed input representations commonly used with perceptrons. Second, each Purkinje neuron receives input from exactly one climbing fiber, and each fiber makes extensive synaptic contact with the dendritic tree of its target Purkinje neuron (Eccles, Ito, & Szentágothai, 1967; Palay & Chan-Palay, 2012). The most effective training methods for artificial neural networks rely on supervised learning algorithms that implement some form of gradient descent (e.g., backpropagation), which require the system to have fine-grained

access to errors that occur at every synapse. The one-to-one correspondence between Purkinje neurons and climbing fibers may be a biologically plausible way of projecting these errors into the cerebellum.

Later physiological discoveries also fall roughly in line with this classic view of the cerebellum. For instance, in Purkinje neurons, the shape of the spike evoked by activation of parallel fibers (i.e., "simple spike") is different from the shape of the spike evoked by inferior olive activation (i.e., "complex spike"). Simple spikes encode parameters of movement such as trajectory, velocity, and acceleration (Gomi *et al.*, 1998; Shidara *et al.*, 1993), whereas complex spikes encode errors in movement (Kitazawa, Kimura, & Yin, 1998; Kobayashi *et al.*, 1998), which is compatible with their involvement in a learning process. Furthermore, the granule cell/Purkinje cell synapse is highly plastic (e.g., it exhibits LTP and LTD both presynaptically and postsynaptically), and climbing fiber signals can control the direction of plasticity (e.g., LTP vs. LTD) at granule cell/Purkinje cell synapses (Coesmans *et al.*, 2004; Lev-Ram *et al.*, 2003). Much is known about the intracellular signalling cascades that drive this plasticity (van Woerden *et al.*, 2009), but the details are beyond the scope of this chapter.

The mechanisms of synaptic plasticity at parallel fiber/Purkinje cell synapses do not fall neatly into the network architectures assumed by two-factor and three-factor learning rules. The two-factor learning rule describes synaptic plasticity when only two neurons are connected (i.e., a presynaptic neuron and a postsynaptic neuron), and the three-factor learning rule describes plasticity when a presynaptic neuron and a dopaminergic input converge on a postsynaptic neuron. In contrast, plasticity at parallel fiber/Purkinje neuron synapses is determined by the convergence of parallel fibers and climbing fibers – both of which are excitatory glutamatergic projections – onto Purkinje neurons. Thus, synaptic plasticity at parallel fiber/Purkinje cell synapses follows its own unique learning rule. In particular, LTD is induced with (1) strong presynaptic activation from input 1, (2) strong presynaptic activation from input 2, and (3) strong postsynaptic activation. In contrast, LTP is induced with (1) weak presynaptic activation from input 1, (2) weak or absent activation from presynaptic input 2, and (3) weak postsynaptic activation. A further difference is that, in the two-factor learning rule, strong presynaptic activation (i.e., above the threshold for NMDA receptor activation) leads to LTP, and weak presynaptic activation leads to LTD. At parallel-fiber/Purkinje neuron synapses, these roles are reversed: weak activation of presynaptic Purkinje neurons leads to LTP, and strong activation leads to LTD.

Finally, we now know that there is synaptic plasticity at a multitude of synapses within the cerebellar circuit beyond those postulated by the classic model (e.g., between mossy fibers, between Purkinje cells and deep cerebellar nuclei, between various interneuron types, etc.), and we understand much of the cellular and molecular mechanisms at play. A complete review of these forms of plasticity and their mechanisms is outside the scope of this chapter, but see D'Angelo (2014) for a review.

Example Models of Supervised Learning in the Cerebellum

Classic models view the cerebellum as a neural implementation of a supervised-learning machine (Albus, 1971; Ito, 1984; Marr, 1969). In this conception, sensory input signals are carried by the mossy fibers, transformed into a more expansive basis set by the greatly divergent projections to the granule neurons, and ultimately transformed into the output signal by the granule neuron projections to Purkinje neurons. The ω parameters of Equation (4.29) denote the synaptic strengths of the connections between granule and Purkinje neurons in this system. Climbing fibers from the inferior olive are thought to provide a supervised error or teaching signal that dictates plasticity at the granule neuron/Purkinje neuron synapse.

Owing largely to the homogeneity of anatomical circuitry across the cerebellum, this basic model has been proposed to apply to essentially every domain of cognition and action (Schmahmann *et al.*, 2019). However, likely because of the cerebellum's early association with motor function, the most clearly developed class of cerebellar-based supervised-learning models include models of motor planning and motor control – especially for arm-reaching movements (Schweighofer, Arbib, & Kawato, 1998; Schweighofer *et al.*, 1998; Wolpert, Miall, & Kawato, 1998). In this case, all signals in Equation (4.29) are considered to vary continuously in time, with output signals $y_j(t)$ conceived of as motor commands (i.e., muscle activation or joint torques), and input signals $x_i(t)$ conceived of as desired trajectories (i.e., position, velocity, and acceleration). In addition, the $\omega_{i,j}$ parameters represent synaptic weights between the granule cell and Purkinje cell layer, and the inferior olive is hypothesized to transmit a supervised error signal (actual trajectory minus desired trajectory).

4.5.5 Models of Human Learning that Include Multiple Forms of Plasticity

After long periods of practice, almost any behavior can be executed quickly, accurately, and with little or no conscious deliberation. At this point, we say that the behavior has become automatic. A strong case can be made that most behaviors performed by adults are automatic. When we sit in a chair, pick up a cup of coffee, or swerve to avoid a pothole, our actions are almost always automatic.

Automaticity could be viewed as the asymptotic state of learning. Ashby, Ennis, and Spiering (2007) proposed that skills learned procedurally are mediated entirely within the cortex after they become automatized, and that the development of automaticity is associated with a gradual transfer of control from the striatum to cortical–cortical projections from the relevant sensory areas directly to the premotor areas that initiate the behavior. So in Figure 4.5, the cortical–cortical projections from the visual cortex to the supplementary motor area eventually mediate the expression of automatic behaviors without any assistance from the subcortical loops through the basal ganglia. Therefore, according to this account, a critical function of the basal ganglia is to train purely cortical representations of automatic behaviors. Kovacs *et al.* (2021) proposed a similar account of how

rule-guided behaviors are automatized in which the prefrontal cortex trains the cortical circuits that implement the automatic behaviors.

The Ashby *et al.* (2007) model was motivated by the observation that because cortical synaptic plasticity follows two-factor learning rules, the purely cortical circuits are incapable of learning any behavior that requires trial-by-trial feedback. Such behaviors require the three-factor plasticity of the basal ganglia. Ashby *et al.* (2007) proposed that the basal ganglia use DA-mediated three-factor learning (i.e., at cortical–striatal synapses) to gradually activate the correct postsynaptic targets in the supplementary motor area, which thereby enables two-factor plasticity at cortical–cortical synapses to learn the correct associations (i.e., because there will be more postsynaptic activation at the correct synapses than at synapses leading to incorrect responses). As a result, in the full version of the Figure 4.5 model, plasticity at cortical–striatal synapses is modeled via three-factor learning rules [as in Equation (4.35)], whereas plasticity at cortical–cortical synapses is modeled via two-factor learning rules [as in Equation (4.33)].

This model accounts for many results that are problematic for other theories of automaticity. For example, it correctly predicts that people with Parkinson's disease, who have DA reductions and striatal dysfunction, are impaired in initial procedural learning (Soliveri *et al.*, 1997; Thomas-Ollivier *et al.*, 1999), but relatively normal in producing automatic skills (Asmus *et al.*, 2008). It also correctly predicts that blocking all striatal output to cortical motor and premotor targets does not disrupt the ability of monkeys to fluidly produce an overlearned motor sequence (Desmurget & Turner, 2010). Similarly, a neuroimaging study reported that activation in the putamen was correlated with performance of a procedural skill early in training but not after automaticity developed (Waldschmidt & Ashby, 2011). Instead, automatic performance was only correlated with activity in cortical areas (i.e., presupplementary motor area and supplementary motor area).

4.6 Empirical Testing

Of course, any psychological theory or model must eventually be tested against empirical data. In the case of learning models, this is especially challenging because, by definition, learning data are nonstationary. In fact, in some cases, the human learner could be in a different state on every trial of the experimental task. If so, then accurate estimation of that state is virtually impossible. In other words, learning data often provide, at best, a highly noisy sample of the learner's true state. As a result, model mimicry is perhaps a greater problem with models of learning than with models of other types of psychological phenomena – that is, learning data are often noisy enough that a less valid model could be statistically indistinguishable from a more valid model, based on goodness-of-fit alone. For these reasons, some extra steps are often needed to test models of learning.

One advantage of building models in which learning is mediated by the synaptic plasticity algorithms described in the previous sections, is that because of their

biological constraints, such models tend to be mathematically rigid (Ashby, 2018). In other words, they tend to make a narrow set of predictions, regardless of how their free parameters are set. Because of this, in many cases, parameter-free *a priori* predictions are possible. For example, any model that assumes learning is based on DA-mediated synaptic plasticity that mimics reinforcement-learning algorithms must predict that omitting trial-by-trial feedback or even delaying feedback by just a few seconds should have devastating effects on learning.

Even if a model does not make *a priori* predictions in a given task, it may predict only a limited set of possible outcomes. If one of those outcomes is observed in an experiment, then a model predicting that this is one of the few outcomes possible should be favored over a model that can account for a wider variety of possible outcomes by manipulating free parameters in a *post hoc* manner. The method of parameter-space partitioning was designed to address this issue (Pitt *et al.*, 2006). In particular, parameter-space partitioning estimates the volume of parameter space throughout which a model is consistent with a certain qualitative pattern of data. A parameter-space partitioning analysis is valuable with all kinds of modeling, but especially so with learning models because of the challenges their nonstationary nature presents to standard goodness-of-fit testing.

Other good model-fitting practices are also recommended. For example, the models should be validated by simulating data under a variety of different parameter settings and then investigating under what conditions the generating parameter values can be recovered during the parameter estimation process.

When learning models are fit to behavioral data, the most common choice is to fit them to some form of empirical learning curve – most often a forward-learning curve, which plots proportion correct against trial or block number. As with all modeling, the most effective tests compare the fit of the model under investigation to some other established model from the literature. In the case of forward learning curves, a good choice for comparison is the exponential learning curve

$$P_n = P_\infty - (P_\infty - P_0)\, \mathrm{e}^{-\lambda n}, \tag{4.43}$$

where P_n is the probability correct on trial n, P_∞ and P_0 are asymptotic and initial accuracy, respectively, and λ is the learning rate. This model was proposed more than 100 years ago (Thurstone, 1919), and remains popular today (e.g., Heathcote, Brown, & Mewhort, 2000; Leibowitz *et al.*, 2010). As an example of how this model might be used, Cantwell *et al.* (2017) compared the fits of the exponential model and a biologically detailed model that assumes learning in procedural-memory-mediated tasks depends on three-factor plasticity (i.e., the model described in Figure 4.5) to learning curves from two separate experiments. In both cases, the biologically detailed model fit better than the exponential model.

Different learning strategies can produce qualitatively different learning curves. Procedural learning and instrumental conditioning predict incremental learning and gradual learning curves. In contrast, rule-guided learning predicts discrete and abrupt jumps in accuracy as the learner switches rules trial-by-trial. In many tasks, incorrect rules cause accuracy to be near chance, whereas the correct rule

predicts perfect accuracy. In these cases, rule-learning strategies predict all-or-none learning curves.

Although incremental and all-or-none learning curves might seem easy to distinguish empirically, it has long been known that these differences can be obscured if the data are averaged across learners (Estes, 1956, 1964). In fact, it is well documented that averaging can change the psychological structure of many different types of data (Ashby, Maddox, & Lee, 1994; Maddox, 1999). As a result, averaging is typically inappropriate when testing models of how individuals learn. For example, if every learner's accuracy jumps from 50% to 100% correct on one trial, but the trial on which this jump occurs varies across participants, then the resulting averaged learning curve will be incremental – not all-or-none (Estes, 1956). The top panel of Figure 4.7 illustrates this phenomenon. This panel shows the traditional (forward) learning curve (i.e., mean accuracy across all participants on every trial) for 1,000 simulated participants who each display all-or-none learning. Specifically, each participant responds randomly with a probability correct of 0.5 until the correct strategy is discovered on some random trial (between 5 and 85), after which they respond perfectly. Note that the all-or-none nature of learning is completely obscured by the averaging process.

Hayes (1953) proposed the backward learning curve as a solution to this problem. Backward learning curves are most effective at discriminating between incremental and all-or-none learning in experiments where perfect accuracy is possible. The first step is to define a learning criterion, which is conservative enough to rule out guessing or partial learning. For example, consider a two-alternative task, like the one illustrated in Figure 4.7, in which the probability correct by guessing is 0.5 on each trial. Then a criterion of 10 consecutive correct responses is possible by guessing with a probability of less than 0.001. A backward learning curve can only be estimated for participants who reach the criterion, so the second step is to separate participants who reached the criterion from those who did not. The most common analysis for nonlearners is to compare the proportion of nonlearners across conditions. The remaining steps proceed for all participants who reached the criterion. Step 3 is to identify for each learner the trial number of the first correct response in the sequence of 10 correct responses that ended the learning phase. Let N_i denote this trial number for learner i. Then note that the response on trial N_i and the ensuing 9 trials were all correct. But also note that the response on the immediately preceding trial (i.e., trial $N_i - 1$) was necessarily an error. Step 4 is to renumber all the trial numbers so that trial N_i becomes trial 1 for every participant. Thus, for every participant, trials 1–10 are all correct responses and trial 0 is an error. The final step is to estimate a learning curve by averaging across learners. The bottom panel of Figure 4.7 shows the backward learning curve that results from this reanalysis of the data plotted in Figure 4.7a.

Because of our renumbering system, the mean accuracy for trials 1–10 will be 100% correct, and the mean accuracy for trial 0 will be 0% correct. Thus, if every learner shows a dramatic one-trial jump in accuracy, then the averaged accuracy on trial -1 should be low, even if the jump occurred on a different trial number for

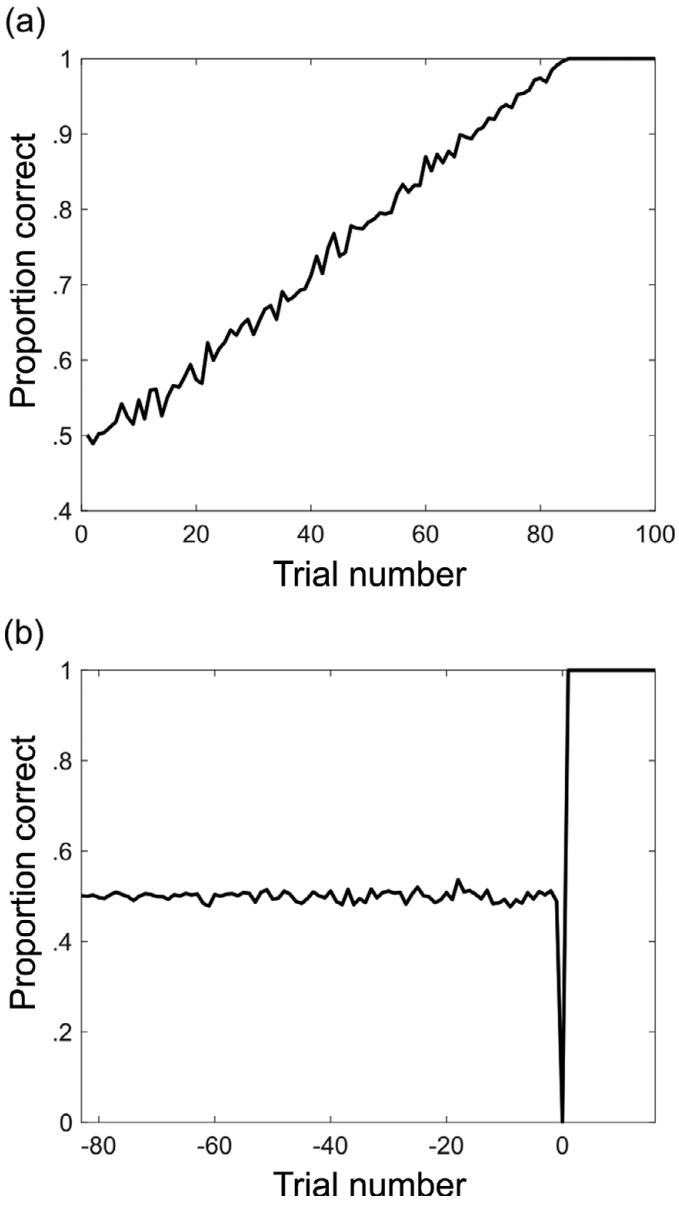


Figure 4.7 *(a) Forward learning curve, which plots mean proportion correct on each trial for 1,000 simulated participants who are all characterized by one-trial learning in which accuracy jumps from 0.5 to 1 on one trial, but who all make this jump on a different random trial. (b) Backward learning curve of the same data.*

every participant (according to the original numbering system). In the Figure 4.7 example, all participants had perfect all-or-none, one-trial learning and note that the mean accuracy for all trials preceding trial 0 is at chance (i.e., 0.5). In contrast, if participants incrementally improve their accuracy then the averaged accuracy

on trial −1 should be significantly higher than chance. So if one is interested in discriminating between strategies that predict incremental learning and strategies that predict all-or-none learning, then backward learning curves should be used rather than the more traditional forward learning curves.

Backward learning curves are more problematic in tasks where most participants do not achieve perfect accuracy, because in these cases, it is usually impossible to define a learning criterion that ensures learning has terminated. Even so, if estimated with care, backward learning curves can be useful even in these more ambiguous cases (J. D. Smith & Ell, 2015).

4.7 Conclusions

Mathematical models of human learning have progressed enormously during the last century. After an initial period of intense activity that dominated experimental psychology during the first half of the twentieth century, the field entered a lull that lasted for several decades. As we have described, several neuroscience breakthroughs reinvigorated the study of learning and the subsequent progress has been dramatic. Even so, the study of learning has not recaptured its formally prominent place within experimental psychology. For example, none of the leading textbooks on cognitive neuroscience currently include any chapters on learning. Learning is a fundamental component of the human experience, and we believe that the recent progress described in this chapter should re-establish the foundational role of learning, not only in mathematical psychology, but more generally within the cognitive sciences.

4.8 Related Literature

Many articles and texts review mathematical learning theory as it existed during the early years of mathematical psychology, including Atkinson, Bower, and Crothers (1965), Bush and Estes (1959), and Laming (1973). No recent texts provide a similar comprehensive coverage. Even so, there are a variety of more specialized recent reviews. In the case of machine learning, the classic text on reinforcement learning is Sutton and Barto (1998), whereas Neal (2012) covers Bayesian approaches. A number of computational neuroscience reviews include sections on learning, including Dayan and Abbott (2001) and Ashby (2018). For a review of the neurobiological foundations of learning (e.g., synaptic plasticity), see Rudy (2020).

Acknowledgments

We thank Sebastien Hélie and Michael Wenger for their helpful comments in the preparation of this chapter.

References

Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences*, *10*(1–2), 25–61.

Alpaydin, E. (2020). *Introduction to machine learning*. Cambridge, MA: MIT Press.

Apicella, P., Legallet, E., & Trouche, E. (1997). Responses of tonically discharging neurons in the monkey striatum to primary rewards delivered during different behavioral states. *Experimental Brain Research*, *116*(3), 456–466.

Ashby, F. G. (2018). Computational cognitive neuroscience. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology, Volume 2* (pp. 223–270). New York: Cambridge University Press.

Ashby, F. G. (2019). *Statistical analysis of fMRI data*, 2nd ed. Cambridge, MA: MIT Press.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Crossley, M. J. (2011). A computational model of how cholinergic interneurons protect striatal-dependent learning. *Journal of Cognitive Neuroscience*, *23*(6), 1549–1566.

Ashby, F. G., & Ennis, J. M. (2006). The role of the basal ganglia in category learning. *Psychology of Learning and Motivation*, *46*, 1–36.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632–656.

Ashby, F. G., & Maddox, W. T. (1998). Stimulus categorization. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making* (pp. 251–301). San Diego, CA: Academic Press.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3), 144–151.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Science*, *2*, 83–89.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science*, 2nd ed. (pp. 157–188). New York: Elsevier.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363–378.

Asmus, F., Huber, H., Gasser, T., & Schöls, L. (2008). Kick and rush: Paradoxical kinesia in Parkinson disease. *Neurology*, *71*(9), 695.

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1965). *Introduction to mathematical learning theory*. New York: Wiley.

Baddeley, R. J., Ingram, H. A., & Miall, R. C. (2003). System identification applied to a visuomotor task: Near-optimal human performance in a noisy changing task. *The Journal of Neuroscience*, *23*(7), 3066–3075.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141.

Bear, M., & Linden, D. (2001). The mechanisms and meaning of long-term synaptic depression in the mammalian brain. In W. Cowan, T. Sudhof, & C. Stevens (Eds.), *Synapses* (pp. 455–517). Baltimore, MD: Johns Hopkins University Press.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Bi, G.-Q., & Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*(1), 139–166.

Bland, A. R., & Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. *Frontiers in Neuroscience*, *6*, 85.

Bliss, T. V., & Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of Physiology*, *232*(2), 331–356.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1485), 1655–1670.

Bush, R. R., & Estes, W. K. (1959). *Studies in mathematical learning theory*. Redwood City, CA: Stanford University Press.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413–423.

Caine, E. D., Weingartner, H., Ludlow, C. L., Cudahy, E. A., & Wehry, S. (1981). Qualitative analysis of scopolamine-induced amnesia. *Psychopharmacology*, *74*(1), 74–80.

Calabresi, P., Pisani, A., Mercuri, N. B., & Bernardi, G. (1996). The corticostriatal projection: From synaptic plasticity to dysfunctions of the basal ganglia. *Trends in Neurosciences*, *19*, 19–24.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, *22*(6), 1598–1613.

Cantwell, G., Riesenhuber, M., Roeder, J. L., & Ashby, F. G. (2017). Perceptual category learning and visual processing: An exercise in computational cognitive neuroscience. *Neural Networks*, *89*, 31–38.

Cheng, S., & Sabes, P. N. (2006). Modeling sensorimotor learning with linear dynamical systems. *Neural Computation*, *18*, 760–793.

Coesmans, M., Weber, J. T., De Zeeuw, C. I., & Hansel, C. (2004). Bidirectional parallel fiber plasticity in the cerebellum under climbing fiber control. *Neuron*, *44*(4), 691–700.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2013). Erasing the engram: The unlearning of procedural skills. *Journal of Experimental Psychology: General*, *142*(3), 710–741.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2014). Context-dependent savings in procedural category learning. *Brain & Cognition*, *92*, 1–10.

Crossley, M. J., Horvitz, J. C., Balsam, P. D., & Ashby, F. G. (2016). Expanding the role of striatal cholinergic interneurons and the midbrain dopamine system in appetitive instrumental conditioning. *Journal of Neurophysiology*, *115*, 240–254.

Crow, T. J., & Grove-White, I. G. (1973). An analysis of the learning deficit following hyoscine administration to man. *British Journal of Pharmacology*, *49*(2), 322–327.

Cunningham, H. A. (1989). Aiming error under transformed spatial mappings suggests a structure for visual-motor maps. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 493–506.

D'Angelo, E. (2014). The organization of plasticity in the cerebellar cortex: From synapses to control. In *Progress in brain research* (Vol. 210, pp. 31–58). Amsterdam: Elsevier.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence* (pp. 761–768). Menlo Park, CA: AAAI Press.

Deserno, L., Boehme, R., Mathys, C., Katthagen, T., Kaminski, J., Stephan, K. E., … Schlagenhauf, F. (2020). Volatility estimates increase choice switching and relate to prefrontal activity in schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(2), 173–183.

Desmurget, M., & Turner, R. S. (2010). Motor sequences and the basal ganglia: Kinematics, not habits. *Journal of Neuroscience*, *30*(22), 7685–7690.

Diaconescu, A. O., Mathys, C., Weber, L. A., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634.

Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In *Stevens' handbook of experimental psychology*. New York: Wiley [online].

Donchin, O., Francis, J. T., & Shadmehr, R. (2003). Quantifying generalization from trial-by-trial behavior of adaptive systems that learn with basis functions: Theory and experiments in human motor control. *Journal of Neuroscience*, *23*(27), 9032–9045.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7–8), 961–974.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(6), 732–739.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859.

Ebner, T. J., & Pasalar, S. (2008). Cerebellum predicts the future motor state. *The Cerebellum*, *7*(4), 583–588.

Eccles, J. C., Ito, M., & Szentágothai, J. (1967). *The cerebellum as a neuronal machine*. New York: Springer.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain*. Oxford: Oxford University Press.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*(2), 94–107.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140.

Estes, W. K. (1964). All-or-none processes in learning and retention. *American Psychologist*, *19*(1), 16–25.

Fanselow, M. S., Zelikowsky, M., Perusini, J., Barrera, V. R., & Hersman, S. (2014). Isomorphisms between psychological processes and neural mechanisms: From stimulus elements to genetic markers of activity. *Neurobiology of Learning and Memory*, *108*, 5–13.

Feenstra, M. G., & Botterblom, M. H. (1996). Rapid sampling of extracellular dopamine in the rat prefrontal cortex during food consumption, handling and exposure to novelty. *Brain Research*, *742*(1), 17–24.

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, *32*, 33–55.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234.

Ghoneim, M., & Mewaldt, S. (1975). Effects of diazepam and scopolamine on storage, retrieval and organizational processes in memory. *Psychopharmacologia*, *44*(3), 257–262.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Suppl. 3), 15647–15654.

Gomi, H., Shidara, M., Takemura, A., Inoue, Y., Kawano, K., & Kawato, M. (1998). Temporal firing patterns of Purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkey I. Simple spikes. *Journal of Neurophysiology*, *80*(2), 818–831.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greeno, J. G., & Bjork, R. A. (1973). Mathematical learning theory and the new "mental forestry". *Annual Review of Psychology*, *24*(1), 81–116.

Grossberg, S. (1972). Neural expectation: Cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. *Kybernetik*, *10*(1), 49–57.

Gu, Q. (2003). Contribution of acetylcholine to visual cortex plasticity. *Neurobiology of Learning and Memory*, *80*(3), 291–301.

Gulliksen, H. (1934). A rational equation of the learning curve based on Thorndike's law of effect. *The Journal of General Psychology*, *11*(2), 395–434.

Guthrie, E. R. (1935). *Psychology of learning*. New York: Harper & Row.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1–2), 1–34.

Hayes, K. J. (1953). The backward curve: A method for the study of learning. *Psychological Review*, *60*(4), 269–275.

Heathcote, A., Brown, S., & Mewhort, D. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304–309.

Houk, J., Adams, J., & Barto, A. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Hull, C. (1943). *Principles of behavior*. New York: Appleton-Century-Crofts.

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*(2), 519–530.

Ito, M. (1984). *The cerebellum and neural control*. New York: Raven Press Books.

Ito, M., Sakurai, M., & Tongroach, P. (1982). Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *The Journal of Physiology*, *324*(1), 113–134.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569-1572.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*(4–6), 535–547.

Kemp, N., & Bashir, Z. I. (2001). Long-term depression: A cascade of induction and expression mechanisms. *Progress in Neurobiology*, *65*(4), 339–365.

Kennedy, A. (2019). Learning with naturalistic odor representations in a dynamic model of the Drosophila olfactory system. *bioRxiv*, 783191.

Kitazawa, S., Kimura, T., & Yin, P.-B. (1998). Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature*, *392*(6675), 494–497.

Kobayashi, Y., Kawano, K., Takemura, A., Inoue, Y., Kitama, T., Gomi, H., & Kawato, M. (1998). Temporal firing patterns of Purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkeys II. Complex spikes. *Journal of Neurophysiology*, *80*(2), 832–848.

Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G. (2021). A neurocomputational theory of how rule-guided behaviors become automatic. *Psychological Review*, *128*(3), 488–508.

Krakauer, J. W., Pine, Z. M., Ghilardi, M.-F., & Ghez, C. (2000). Learning of visuomotor transformations for vectorial planning of reaching trajectories. *The Journal of Neuroscience*, *20*(23), 8916–8924.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–247.

Kruschke, J. K. (2011). *Doing Bayesian data analysis*. Burlington, MA: Academic Press.

Laming, D. R. J. (1973). *Mathematical psychology*. New York: Academic Press.

Lapish, C. C., Kroener, S., Durstewitz, D., Lavin, A., & Seamans, J. K. (2007). The ability of the mesocortical dopamine system to operate in distinct temporal modes. *Psychopharmacology*, *191*(3), 609–625.

Leibowitz, N., Baum, B., Enden, G., & Karniel, A. (2010). The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, *54*(3), 338–340.

Lev-Ram, V., Mehta, S. B., Kleinfeld, D., & Tsien, R. Y. (2003). Reversing cerebellar long-term depression. *Proceedings of the National Academy of Sciences*, *100*(26), 15989–15993.

Lisman, J., Schulman, H., & Cline, H. (2002). The molecular basis of CaMKII function in synaptic and behavioural memory. *Nature Reviews Neuroscience*, *3*(3), 175–190.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, *61*(2), 354–374.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 650–662.

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research*, *74*(2), 219–236.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100–107.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology*, *202*, 437–470.

Martin, S., Grimwood, P., & Morris, R. (2000). Synaptic plasticity and memory: An evaluation of the hypothesis. *Annual Review of Neuroscience*, *23*(1), 649–711.

Martin, T. A., Keating, J. G., Goodkin, H. P., Bastian, A. J., & Thach, W. T. (1996a). Throwing while looking through prisms: I. Focal olivocerebellar lesions impair adaptation. *Brain*, *119*(4), 1183–1198.

Martin, T. A., Keating, J. G., Goodkin, H. P., Bastian, A. J., & Thach, W. T. (1996b). Throwing while looking through prisms: II. Specificity and storage of multiple gaze–throw calibrations. *Brain*, *119*(4), 1199–1211.

Mathys, C. D., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 825.

McCoy, P. A., Huang, H.-S., & Philpot, B. D. (2009). Advances in understanding visual cortex plasticity. *Current Opinion in Neurobiology*, *19*(3), 298–304.

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, *72*(2), 1024–1027.

Mishkin, M., Malamut, B., & Bachevalier, J. (1984). Memories and habits: Two neural systems. In G. Lynch, J. L. McGaugh, & N. M. Weinberger (Eds.), *Neurobiology of human learning and memory* (pp. 65–77). New York: Guilford Press.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. Cambridge, MA: MIT Press.

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 689.

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 855–874.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378.

Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Berlin: Springer Science & Business Media.

Nicoll, R. A. (2017). A brief history of long-term potentiation. *Neuron*, *93*(2), 281–290.

Ninkovic, J., & Bally-Cuif, L. (2006). The zebrafish as a model system for assessing the reinforcing properties of drugs of abuse. *Methods*, *39*(3), 262–274.

O'Reilly, R. C., Munakata, Y., Frank, M., Hazy, T., *et al.* (2012). *Computational cognitive neuroscience*. Mainz: PediaPress.

Ostfeld, A. M., & Aruguete, A. (1962). Central nervous system effects of hyoscine in man. *Journal of Pharmacology and Experimental Therapeutics*, *137*(1), 133–139.

Pakhotin, P., & Bracci, E. (2007). Cholinergic interneurons control the excitatory input to the striatum. *The Journal of Neuroscience*, *27*(2), 391–400.

Palay, S. L., & Chan-Palay, V. (2012). *Cerebellar cortex: Cytology and organization*. Berlin: Springer Science & Business Media.

Paliwal, S., Mosley, P., Breakspear, M., Coyne, T., Silburn, P., Aponte, E., … Klaas, S. (2018). Subjective estimates of uncertainty and volatility during gambling predict impulsivity after subthalamic deep brain stimulation for Parkinson's disease. *BioRxiv*, 477364.

Paliwal, S., Petzschner, F. H., Schmitz, A. K., Tittgemeyer, M., & Stephan, K. E. (2014). A model-based analysis of impulsivity using a slot-machine gambling paradigm. *Frontiers in Human Neuroscience*, *8*, 428.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Translated and edited by G. V. Anrep. London: Oxford University Press.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, *7*(1).

Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, *59*(1), 319–330.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83.

Poldrack, R. A., Clark, J., Pare-Blagoev, E., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550.

Rall, W. (1967). Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input. *Journal of Neurophysiology*, *30*(5), 1138–1168.

Ramnani, N. (2006). The primate cortico-cerebellar system: Anatomy and function. *Nature Reviews Neuroscience*, *7*(7), 511–522.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163–178.

Redding, G. M., Rossetti, Y., & Wallace, B. (2005). Applications of prism adaptation: A tutorial in theory and method. *Neuroscience & Biobehavioral Reviews*, *29*(3), 431–444.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Reynolds, J. N. J., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, *15*, 507–521.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92*(3), 365–372.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rudy, J. W. (2020). *The neurobiology of learning and memory*, 3rd ed. Oxford: Oxford University Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. *Journal of Comparative and Physiological Psychology*, *74*, 192–202.

Sawtell, N. B. (2010). Multimodal integration in granule cells as a basis for associative plasticity and sensory prediction in a cerebellum-like circuit. *Neuron*, *66*(4), 573–584.

Scheidt, R. A., Dingwell, J. B., & Mussa-Ivaldi, F. A. (2001). Learning to move amid uncertainty. *Journal of Neurophysiology*, *86*(2), 971–985.

Schmahmann, J. D., Guell, X., Stoodley, C. J., & Halko, M. A. (2019). The theory and neuroscience of cerebellar cognition. *Annual Review of Neuroscience*, *42*(1), 337–364.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Schweighofer, N., Arbib, M. A., & Kawato, M. (1998). Role of the cerebellum in reaching movements in humans. I. Distributed inverse dynamics control. *European Journal of Neuroscience*, *10*(1), 86–94.

Schweighofer, N., Spoelstra, J., Arbib, M. A., & Kawato, M. (1998). Role of the cerebellum in reaching movements in humans. II. A neural model of the intermediate cerebellum. *European Journal of Neuroscience*, *10*(1), 95–105.

Seamans, J. K., & Robbins, T. W. (2010). Dopamine modulation of the prefrontal cortex and cognitive function. In K. A. Neve (Ed.), *The dopamine receptors*, 2nd ed. (pp. 373–398). New York: Springer.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shadmehr, R., & Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Experimental Brain Research*, *185*(3), 359–381.

Shidara, M., Kawano, K., Gomi, H., & Kawato, M. (1993). Inverse-dynamics model eye movement control by Purkinje cells in the cerebellum. *Nature*, *365*(6441), 50–52.

Sjöström, P. J., Rancz, E. A., Roth, A., & Häusser, M. (2008). Dendritic excitability and synaptic plasticity. *Physiological Reviews*, *88*(2), 769–840.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

Smith, J. D., & Ell, S. W. (2015). One giant leap for categorizers: One small step for categorization theory. *PloS One*, *10*(9).

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168.

Soliveri, P., Brown, R. G., Jahanshahi, M., Caraceni, T., & Marsden, C. D. (1997). Learning manual pursuit tracking skills in patients with Parkinson's disease. *Brain*, *120*(8), 1325–1337.

Soto, F. A., & Wasserman, E. A. (2010). Error-driven learning in visual categorization and object recognition: A common-elements model. *Psychological Review*, *117*(2), 349–381.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Takahashi, Y., Schoenbaum, G., & Niv, Y. (2008). Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in Neuroscience*, *2*, 14.

Thomas-Ollivier, V., Reymann, J., Le Moal, S., Schück, S., Lieury, A., & Allain, H. (1999). Procedural memory in recent-onset Parkinson's disease. *Dementia and Geriatric Cognitive Disorders*, *10*(2), 172–180.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, *39*(1/4), 212–222.

Thoroughman, K. A., & Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. *Nature*, *407*(6805), 742–747.

Thurstone, L. L. (1919). The learning curve equation. *Psychological Monographs*, *26*(3), i–51.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *The Journal of Neuroscience*, *23*(32), 10402–10410.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tulving, E., & Craik, F. I. (2000). *The Oxford handbook of memory*. New York: Oxford University Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Valentin, V. V., Maddox, W. T., & Ashby, F. G. (2014). A computational model of the temporal dynamics of plasticity in procedural learning: Sensitivity to feedback timing. *Frontiers in Psychology*, *5*(643).

van Woerden, G. M., Hoebeek, F. E., Gao, Z., Nagaraja, R. Y., Hoogenraad, C. C., Kushner, S. A., … Elgersma, Y. (2009). βCaMKII controls the direction of plasticity at parallel fiber-Purkinje cell synapses. *Nature Neuroscience*, *12*(7), 823–825.

Von Helmholtz, H. (1925). *Helmholtz's treatise on physiological optics* (Vol. 3). Washington, DC: Optical Society of America.

Waldschmidt, J. G., & Ashby, F. G. (2011). Cortical and striatal contributions to automaticity in information-integration categorization. *NeuroImage*, *56*(3), 1791–1802.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158–177.

Weilnhammer, V. A., Stuke, H., Sterzer, P., & Schmack, K. (2018). The neural correlates of hierarchical predictions for perceptual decisions. *Journal of Neuroscience*, *38*(21), 5008–5021.

Weiss, E. J., & Flanders, M. (2011). Somatosensory comparison during haptic tracing. *Cerebral Cortex*, *21*(2), 425–434.

Welch, R. B. (1986). Adaptation of space perception. *Handbook of perception and human performance*, *1*(24), 2424–2445.

Wickens, J. (1993). *A theory of the striatum*. New York: Pergamon Press.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (Tech. Rep.). Stanford University, California, Stanford Electronics Labs.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review*, *105*(3), 558–584.

Wills, A., Noury, M., Moberly, N. J., & Newport, M. (2006). Formation of category representations. *Memory & Cognition*, *34*(1), 17–27.

Wilson, C. J. (1995). The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 29–50). Cambridge, MA: MIT Press.

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Computational Biology*, *9*(7).

Wolpert, D. M., Miall, R., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Sciences*, *2*(9), 338–347.

Worthy, D. A., Markman, A. B., & Maddox, W. T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition*, *81*(2), 283–293.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345*(6204), 1616–1620.

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A., & Poo, M.-M. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, *395*(6697), 37–44.

5 Formal Models of Memory Based on Temporally-Varying Representations

Marc W. Howard

Human babies, while adorable, are remarkably incompetent. They know essentially no facts about the world, are unable to perform any but the simplest motor actions, and perform very poorly on behavioral assays of memory. Memory researchers evaluate memory in adults with a variety of behavioral paradigms, such as cued recall, in which the participant is given a series of pairs (e.g., ABSENCE–HOLLOW, PUPIL–RIVER, CAMPAIGN–HELMET). The participants' task is to produce the correct associate when given a cue word. For instance, after being probed with PUPIL, the correct response is RIVER. After being presented with a list of words for a cued-recall test, a human baby is more likely to emit curdled milk than a correct response. Over the course of a lifetime, normally developing humans learn many

facts about their world, acquire complicated motor skills, and can bring to mind vivid recollections of many events from their lives. Because all of these abilities must be learned, they can be understood as forms of memory.

Viewed in this light, the task of a memory theorist seems daunting. How can one possibly construct a theory that can make sense of the ability to recall that Paris is the capital of France, the ability to ride a bike without falling over, *and* the ability to vividly remember a birthday party well enough to bring a smile to one's face after decades? The strategy taken by cognitive neuroscientists in the latter part of the twentieth century (and continuing to the present day) is to carve up the set of abilities and skills that differentiate a baby from an adult into different "kinds" of memory, each associated with distinct parts of the brain. For instance, many memory researchers would say that retrieving facts about the world depends on semantic memory, being able to ride a bicycle is a consequence of implicit memory, and vivid recollection of specific events from one's life relies on episodic memory. This strategy of dividing learning and memory phenomena into different "kinds of memory" has been extremely productive. However, throughout the history of psychology, there has been an urge towards developing unified theories of learning and memory.

5.0.1 Associations in the Mind and Brain

Radical behaviorists (most famously B. F. Skinner) attempted to understand the rich repertoire of memory phenomena as special cases of stimulus–response associations. Pavlov's dogs learned to associate the sound of a bell with the delivery of food, so that the sound of the bell by itself leads to an overt response (salivation). Experimentalists learned that animals (in particular rats and pigeons) can be trained to perform complex sequences of behaviors in response to appropriate training experiences. According to behaviorists' conceptions of learning, even complex behaviors could be described as complex chains of simple associations.

Mathematical psychologists have developed formal models of association to provide quantitative models of behavior in a variety of experimental paradigms. Early work focused on animal conditioning experiments. In this case the behavioral measure is typically a scalar value that describes the probability or magnitude of a conditioned response; for instance, the amount of saliva produced by Pavlov's dog (or, more typically, the proportion of time the animal spends freezing in a fear conditioning experiment). But later work also applied similar ideas to memory experiments with humans using lists of words as stimuli. In the cued-recall task described above, it is straightforward to write down a model that constructs simple associations between neural representations of the words (e.g., associate ABSENCE to HOLLOW) such that probing the memory with the stimulus ABSENCE causes a pattern like HOLLOW to be produced as an output. These models can produce many distinct responses in response to many different cues.

Associations can be understood neurally as a consequence of changes in the connection strength between neurons. The mammalian brain contains a great

number of specialized cells called neurons. Neurons are known to communicate information between one another by means of their electrical activity. The connections between individual neurons are referred to as synapses. The strength of synapses can be modified by experience. These facts are sufficient to write down a very crude neural model of Pavlovian conditioning. If one identifies the set of neurons that changes its firing in response to the sound of the bell, and the set of neurons responsible for salivation, one could in principle understand the association learned by Pavlov's dog as an increase in the strength of the synapses connecting the "bell" neurons to the "drool" neurons. These assumptions can be formalized in tractable mathematical models that are (at least) neurally reasonable. Extending this idea to models of more elaborate tasks, such as human cued recall, requires mapping each of the stimuli that will be part of the experiment (i.e., each of the words in the list) to a pattern of activation over neurons. This is typically done by mapping each word to a vector in a space of neurons. In this case, the synapses between the neurons can be understood as a matrix. With appropriate assumptions, many results can be derived and a particular set of assumptions can be compared to behavior.

5.0.2 Cognitive Models of Memory

The basic theoretical stance of behaviorism is that we should construct psychological theory without reference to the internal state of the organism. This approach is difficult to reconcile with many human laboratory memory tasks. For instance, a radical behaviorist model of the free-recall task is untenable. In free recall, participants are presented with a sequential experience (e.g., a list of words) and later asked to verbally report their memory for the experience. What is the "cue" in free recall? Participants can report many different experiences and can report on different aspects of their experience. It is difficult to make sense of these phenomena without simply assuming that the participant has some internal experience of their memory that they then describe.

Cognitive models make a hypothesis about the internal state of the organism and use that hypothesis to predict behavior. Radical behaviorists explicitly eschewed any reference to the internal experience of the behaving organism in the belief that such theorizing was underconstrained and cannot lead to a satisfactory scientific theory. However, advances in modern neuroscience have made this concern largely obsolete. In principle, cognitive models can simultaneously describe the observable behavior of an organism and neural observables from the brain during performance of that behavior. In this way, cognitive models can be constrained by comparison to activity of neurons in the brain.

A broad class of cognitive models proceed by building simple associations between stimuli mediated by a hypothesized internal state. For instance, short-term memory models hypothesize the existence of a short-term store that holds information about recently presented stimuli. According to one influential approach, associations between stimuli can only be formed among stimuli that are simultaneously

active in the short-term store (Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1980). Another widely used approach assumes the existence of a "temporal context" that mediates associations between items (Polyn, Norman, & Kahana, 2009; Sederberg, Howard, & Kahana, 2008). Temporal-context models assume that the brain maintains a representation at each moment of the recent past. This temporal context changes gradually. When a person remembers a specific instance from their past (like vividly remembering a particular event such as a birthday), this cognitive event is accompanied by a recovery of temporal context. These models make specific neural predictions. Short-term memory models and temporal-context models predict that it ought to be possible to examine the activity of neurons in the brain (using electrodes or noninvasive methods such as EEG or fMRI) and decode the content of recent experiences. Cognitive models of this class are introduced in Section 5.2.

5.0.3 Beyond Associations: Representing Temporal Relationships in the Mind and Brain

Although associations have been an extremely productive idea in the mathematical psychology of memory, there is no question that simple associations as understood by behaviorists are insufficient to describe the richness of human memory. Associations that can be described by a scalar value are extremely limited. If the association between stimulus X and stimulus Y is some specific number, say 2.38, and the association between X and Z is 0.35, we can say that the $X \rightarrow Y$ association is stronger than the $X \rightarrow Z$ association. Operationally, if we probe memory with X, memory returns "more" Y than Z. However, human memory can learn and express many different *kinds* of relationships. For instance, X might be 2 m to the east of Y, or X might be a member of the category Z, or Y and Z might be married to one another. In order to express these kinds of relationships, a richer formalism is required.

The mammalian brain contains neurons that can express metric relationships between stimuli. For instance, consider neurons referred to as "time cells" in the rodent hippocampus during performance of a behavioral task (Eichenbaum, 2017). After presentation of a stimulus (e.g., ringing a bell), these time cells fire in a sequence such that each neuron fires for a circumscribed period of time (see Figure 5.6 later). Because the sequence is reliable across different presentations of the same stimulus, it is possible to look at which time cell is firing and decode how far in the past the triggering stimulus was experienced. As we will see, the information about the time in the past at which the bell was presented written across this population of neurons can be used to learn temporal *relationships* between the presentation of the bell and other stimuli. This class of models has been used to develop cognitive models of relatively complex behavioral tasks and at the same time the properties of time cells can be evaluated against experiments recording from populations of neurons in mammals. To the extent that this hypothesis is consistent with both behavioral and neurophysiological data,

it makes sense to take the equations seriously. As we will see, the formalism is quite rich, providing an opportunity to do meaningful theoretical work on physical models of memory.

5.0.4 A Brief History of Mathematical Models of Memory

This chapter covers a tiny proportion of the work in mathematical models of human memory. To provide at least pointers to the topics that are missing, and to properly contextualize the topics that are covered, this subsection provides a very concise history of mathematical models of memory.

Descriptive quantitative models of behavior date back to the very beginning of modern memory research. Ebbinghaus conducted early empirical studies of human memory, testing himself on serial recall of nonsense syllables and including quantitative descriptions of many of the phenomena he studied (Ebbinghaus, 1885/1913). For instance, Ebbinghaus introduced the power law of forgetting to describe his findings relating the persistence of memory to the passage of time. In the early part of the twentieth century, radical behaviorism led many researchers to focus on simple stimulus–response associations. Quantitative models of these data attempted to describe observable phenomena with as few assumptions as possible. Hull (1939) provides an excellent example of the spirit of this work, fitting equations to observed empirical relationships.

The 1950s saw the first process models of memory. Process models, in contrast to descriptive models, make hypotheses about internal mechanisms that cause observable behavior. Stimulus-sampling theory (Bush & Mosteller, 1951; Estes, 1950) provides an early example of such a process model. Stimulus-sampling theory introduced a number of ideas that are still extremely influential today (see Section 5.1).

The 1960s and 1970s saw memory research divide into a set of subfields as the cognitive revolution dramatically changed the kinds of theories that were acceptable in psychology. There were two major developments in mathematical models of memory during this era that had long-lasting effects over the next several decades. First, building on a long tradition of mathematical models of conditioning, the Rescorla–Wagner model (Rescorla & Wagner, 1972) successfully accounted for essentially everything that was known about classical conditioning up to that time. The Rescorla–Wagner model is built on a really simple idea – that change in an association between a cue and a response depends on how well the outcome is predicted. Second, the 1960s saw the development of the first models of short-term memory building on early ideas from Miller (1956). The two-store memory model of Atkinson and Shiffrin (1968) provided a conceptually simple description of an immense amount of data (see Section 5.2). This was also perhaps the first influential mathematical model of memory to make use of computer simulations to test its predictions. These two very different models spawned entire fields of research in psychology and neuroscience that continue to this day.

The Rescorla–Wagner model led directly to reinforcement learning (Sutton & Barto, 1981). Reinforcement learning has been extremely influential in neuroscience, where the connection between these models and the dopamine system in the brain (Schultz, Dayan, & Montague, 1997) has spawned an immense amount of work that continues to the present day (e.g., see Chapter 4 in this volume). Reinforcement learning has also been extremely influential in artificial-intelligence research, including very high-profile papers building models to achieve human-level performance in video games and the game of go (Mnih *et al.*, 2015; Silver *et al.*, 2016).

The Atkinson and Shiffrin (1968) model also led to a great deal of work in psychology and neuroscience. The model coincided with the discovery of patients with brain damage that showed problems with short-term memory but not long-term memory, and vice versa. Baddeley and Hitch (1977) further subdivided the short-term store and mapped these components onto distinct brain circuits. This kind of model – with many components that map onto different parts of the brain – was well suited for posing the kinds of questions that could be answered with early cognitive neuroimaging techniques such as PET and univariate fMRI. Mathematical models of short-term memory continue to be influential in contemporary cognitive neuroscience (see Trutti *et al.*, 2021 for a recent review).

In the 1980s and 1990s, a great deal of attention was focused on a class of mathematical models of memory that were collectively known as distributed memory models. These models focused on human memory experiments, primarily experiments that would be understood today as episodic memory tasks. Models that fall into this class include TODAM (Murdock, 1982), CHARM (Metcalfe, 1985), SAM (Gillund & Shiffrin, 1984), MINERVA-2 (Hintzman, 1987), the matrix model (Humphreys, Bain, & Pike, 1989), and REM (Shiffrin & Steyvers, 1997). Although these models differed in many details, there were some common assumptions. First, they represented studied items as a distributed set of features, building on early work by Anderson (1972, 1973). Section 5.1 also adopts this convention. Second, the distributed memory models were all associative. It was implicit that short-term memory controlled which items and associations were stored in memory. An important conceptual contribution of these models was the introduction of quantitative models for context (see especially Murdock, 1997) that we build on in Section 5.2. The temporal context models discussed in Section 5.2 grew out of this tradition.

The early distributed memory models did not make a connection to neuroscience. In contrast, connectionist models of memory (see Hasselmo & McClelland, 1999 for a review of early work) paid close attention to neuroscience. For the most part, these models did not focus on detailed behavioral data from human memory experiments (but see Hasselmo & Wyble, 1997; Norman & O'Reilly, 2003). Rather, these models focused more on problems, such as amnesia and sleep, that had a clear connection to neural processes. For instance, in one very influential paper, McClelland, McNaughton, and O'Reilly (1995) postulated that behavioral patterns observed in amnesia patients – for instance the ability to remember events

from early in one's life but not more recent events – were attributable to separate memory stores that learned associations with different statistics. Connectionist memory models were developed in parallel with advances in artificial neural networks that are fundamental to contemporary AI.

One very important development in the early part of the twentieth century was that models of conditioning made contact with models of timing behavior. Scalar expectancy theory (Gibbon, 1977) provided an excellent model of behavioral experiments where animals had to use their sense of time to receive a reward (see also Killeen & Fetterman, 1988). Gallistel and Gibbon (2000) constructed a mathematical model out of scalar expectancy theory that described a range of findings from conditioning experiments. The hypothesis was that behavioral associations fundamentally result from learning about the temporal relationships between the stimulus and response. Balsam and Gallistel (2009) provide an elegant overview of this idea. Notably, because timing behavior has the same properties over a range of time scales, models of conditioning built on this assumption can naturally accommodate scale invariance in memory, which is discussed further in Section 5.2.6.

Section 5.3 draws on work over the last decade or so that synthesizes aspects of many of these approaches. The scale-invariant temporal history was originally proposed to address limitations in temporal context models (Shankar & Howard, 2010). As such, it is continuous with the distributed memory models and can be used to build models of similar tasks. At the same time, because neuroscientific considerations place such strong constraints on these models, it is similar in spirit to the connectionist models of memory. Finally, because memory traces are formed using a population that contains information about the time at which events took place, this approach is closely related to (and in actual fact was very much inspired by) work pursuing a close relationship between timing and conditioning.

5.1 "Simple" Associations in the Mind and Brain

In this section we will introduce a formalism to describe mathematical models based on simple associations. We will suppose that learning consists of forming and accessing associations between a set of "items." These items can correspond to words in a cued-recall experiment, in which we attempt to describe the association between two words (e.g., ABSENCE–HOLLOW above). Or we could use the same formalism to describe the association between a tone that serves as a conditioned stimulus and an unconditioned response, such as salivation in the case of Pavlov's dog.

Distributed memory models (DMMs) assume that each item is described by a vector over some high-dimensional space. We will write vectors as lower-case bold letters, $\mathbf{v} = \{v_1, v_2, \ldots, v_n\}^T$, where n is some "large" integer. We can envision the vector as a list of numbers that describes the activity over a large population of neurons. If a particular item $\mathbf{v}$ represents a word, we might understand $\mathbf{v}$ as the "pattern of activity" over a population of neurons that are caused by presentation

of that word. A different word would produce a different pattern of activity. If a particular item corresponds to a response, such as salivation, we might understand **v** as the pattern of activity in a particular population that is necessary for salivation rather than some other response (such as freezing).

5.1.1 Hebbian Learning

As an illustration of the distributed-memory model approach, let us consider a simple model of cued recall. We map all of the words that could possibly be presented in an experiment onto a set of vectors within the same space. We assume further that the overwhelming majority of entries in each vector **v** are zero and the remainder are some small positive number and that the number of entries n is large. Suppose that we randomly choose vectors corresponding to two different words $\mathbf{v}_i$ and $\mathbf{v}_{j\neq i}$. We can take the inner product between any two vectors as a measure of their "similarity." With these assumptions, the inner product of a vector with itself, $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i$ or $\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j$, will tend to be much greater than the inner product between different words $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j$, because the entries of these are not perfectly correlated. We might even suppose that related words (e.g., COUCH and SOFA) correspond to vectors that are more similar to one another than unrelated words (e.g., COUCH and RUTABAGA). To keep the arithmetic simple, let us suppose that we have chosen the entries in the vectors to ensure that the expected value of $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j$ is 1 if $i = j$ and effectively 0 otherwise.

Let us flesh this model out sufficiently to model a simple cued-recall experiment. Let us describe a list of word pairs by denoting the cue of the pair presented at time t with a vector $\mathbf{f}_t$ and the response member of the pair with a vector $\mathbf{g}_t$. So, if we had a list of two pairs, ABSENCE–HOLLOW and PUPIL–RIVER, we would refer to the vector corresponding to ABSENCE as $\mathbf{f}_1$, the vector corresponding to HOLLOW as $\mathbf{g}_1$, the vector corresponding to PUPIL as $\mathbf{f}_2$ and RIVER as $\mathbf{g}_2$.

Now, we can model associations between the words as an outer product matrix between the vectors corresponding to the cue and response of each pair. Let us assume that the matrix **M** is initialized as an $n \times n$ matrix of zeros before the list. Then as each item is presented, **M** is updated as

$$\Delta\mathbf{M}_t = \mathbf{g}_t\mathbf{f}_t^{\mathrm{T}} \tag{5.1}$$

so that after learning the entire list:

$$\mathbf{M} = \sum_t \mathbf{g}_t\mathbf{f}_t^{\mathrm{T}}, \tag{5.2}$$

where the sum is over all of the pairs presented in the experiment.

To understand the role of the outer product, let us imagine we have a one-pair list so that $\mathbf{M} = \mathbf{g}\mathbf{f}^{\mathrm{T}}$ (Figure 5.1). Any particular entry $M_{ij} = g_i f_j$ gives the product of the activity in "neuron i" in pattern **g** and "neuron j" in pattern **f**. The product is nonzero if both g_i and f_j are both nonzero. The anatomical structure that connects the axon of one neuron to the dendrite of another is referred to as a synapse. These

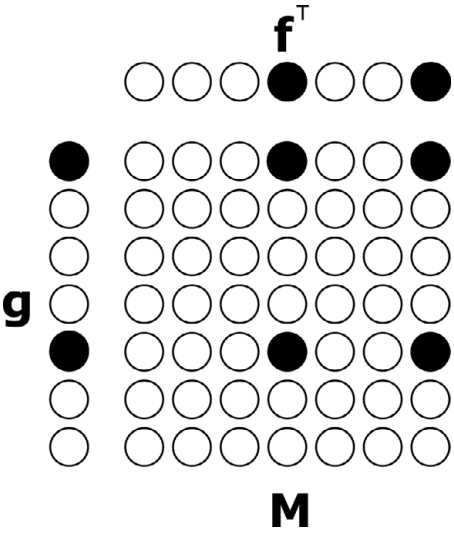


Figure 5.1 *Graphic illustration of the equation* $\mathbf{M} = \mathbf{g}\mathbf{f}^{\mathrm{T}}$. *Here,* $\mathbf{f}$ *is a vector that is zero except for entries* 4 *and* 7*;* $\mathbf{g}$ *is a vector that is zero except for entries* 1 *and* 5. *The outer product matrix* $\mathbf{M}$ *is zero except for entries where both* $\mathbf{f}$ *and* $\mathbf{g}$ *had nonzero values. Probing as* $\mathbf{M}\mathbf{f}$ *gives back* $\mathbf{g}$ *multiplied by the squared length of* $\mathbf{f}$.

connections can be strengthened or weakened based on the activity of the pre- and postsynaptic neurons through a variety of molecular processes. Hebbian learning (originally proposed by Donald Hebb in 1948) is a learning rule in which synapses are strengthened if both the pre- and postsynaptic neurons are active at the same time (see Chapter 4 in this volume for more details). Informally, Hebbian learning is often summarized by the slogan "neurons that fire together, wire together." Hebbian learning has been demonstrated experimentally in a number of brain regions and a number of species.

To understand why this is referred to as an association, let us probe $\mathbf{M}$ with a probe word, which we denote as $\mathbf{f}_p$. Then we find that $\mathbf{M}\mathbf{f}_p = \left(\mathbf{g}\mathbf{f}^{\mathrm{T}}\right)\mathbf{f}_p = \mathbf{g}\left(\mathbf{f}^{\mathrm{T}}\mathbf{f}_p\right)$. That is, probing $\mathbf{M}$ with a probe vector $\mathbf{f}_p$ returns $\mathbf{g}$ weighted by the similarity between the probe vector and the studied cue vector. If the probe vector $\mathbf{f}_p$ is the same as the studied cue $\mathbf{f}$, the output is $\mathbf{g}$ multiplied by a large number. If $\mathbf{f}_p$ is not the same as $\mathbf{f}$, the output is $\mathbf{g}$ mutiplied by a small number. Returning to the situation where there are many pairs in the list, we find (exploiting the linearity of matrix addition and commutativity of multiplication by a scalar)

$$\mathbf{M}\mathbf{f}_p = \left[\sum_t \mathbf{g}_t\mathbf{f}_t^{\mathrm{T}}\right]\mathbf{f}_p \tag{5.3}$$

$$= \sum_t \left(\mathbf{f}_t^{\mathrm{T}}\mathbf{f}_p\right)\mathbf{g}_t. \tag{5.4}$$

That is, after probing memory with a specific word $\mathbf{f}_p$, the output is the vector sum of the response words $\mathbf{g}_t$ weighted by the similarity of the probe word to the cue that was paired with that response. Because the similarity of the probe words to themselves is much greater than between different words, this sum gives a large number for the appropriate response and much smaller numers for the other possible responses. If one probes $\mathbf{M}$ with $\mathbf{f}_{\text{ABSENCE}}$, the output is "mostly" $\mathbf{g}_{\text{HOLLOW}}$; if one probes with $\mathbf{f}_{\text{PUPIL}}$, the output is mostly $\mathbf{g}_{\text{RIVER}}$. By adding assumptions that map the output of the associative memory onto a probability of successfully recalling the appropriate response, one can construct relatively elaborate models of behavior.

If each component of $\mathbf{f}$ and $\mathbf{g}$ can be thought of as a neuron, then each entry in $\mathbf{M}$ can be understood as a synapse. The entire matrix $\mathbf{M}$ can thus be understood as the set of synapses connecting the two populations. The outer product learning rule in Equation (5.1) can thus be understood as a simple hypothesis for how populations of neurons can store information via Hebbian learning. Although this is undoubtedly a grotesque oversimplification of what happens in the brain, this framework is sufficiently simple that one can write out tractable models of behavioral experiments.

To actually compare this model to behavioral data, it's necessary to specify some means to map the strength of the association onto behavioral observables, for instance probability of recall. Having said that, this simple Hebbian mechanism responds appropriately to many experimental manipulations in a sensible way. For instance, suppose that some pairs in the list are repeated. Adding additional terms with the same vectors to Equation (5.2) results in a stronger association between those items (this follows from linearity).[1] Similarly, one can compare recall of a particular pair in lists of various lengths. Examining Equation (5.4), we see that the effect of including additional pairs is to add noise to the output of memory. That is, after probing with $\mathbf{f}_i$, the output of memory is $\mathbf{g}_j$ times a big number plus all of the other items in the list weighted by small numbers. As one adds pairs to the list, this second component grows more important, acting like background noise for retrieval of the target response. Similarly, one could imagine that attention fluctuates from moment-to-moment and model that by multiplying Equation (5.1) by a factor that estimates the current amount of attention. Distributed memory models pursued questions along these lines and carefully compared the results to behavioral experiments.

5.1.2 Forgetting

The Hebbian outer-product model sketched above has several problems, many of which are addressed by subsequent work described in the remainder of this chapter. Here we discuss ways to enable the model to *forget*. We discuss two approaches

1 One can easily construct a similar argument for the effect of increasing the study time for some of the pairs in the list.

to forgetting. Perhaps the most obvious way to implement forgetting is to allow the weights to decrease in amplitude. A less obvious way to implement forgetting is to assume that the cue itself is not constant over time. That is, although an experimenter may take care to present the word ABSENCE several times in the same font, in the same location of the screen, for precisely the same duration of time, this does not ensure that this stimulus activates the same set of neurons in the brain on each presentation. There are many other possible approaches to forgetting and different mechanisms may contribute differentially to forgetting in different experimental paradigms. This chapter focuses on these two mechanisms for forgetting because they lend themselves to concise mathematical descriptions and are conceptually distinct from one another.

Forgetting via Changes in the Weight Matrix

One simple way to augment Equation (5.1) to enable forgetting is to allow the weights to decay exponentially as a function of time:

$$\mathbf{M}_{t+1} = \rho \mathbf{M}_t + \mathbf{g}_t \mathbf{f}_t^{\mathrm{T}}, \tag{5.5}$$

where $0 < \rho < 1$. Each additional time step results in an additional power of ρ, so that the output caused by a memory probe decreases the longer it has been available in memory. After studying L items, we find

$$\mathbf{M}_L \mathbf{f}_p = \sum_t \rho^{L-t} \left(\mathbf{f}_t^{\mathrm{T}} \mathbf{f}_p \right) \mathbf{g}_t. \tag{5.6}$$

The last term shows that the strength of the association stored in $\mathbf{M}$ decays exponentially as a function of how far in the past the association was learned.

One of the longest-standing questions in memory research is whether we forget over time due to the passage of time per se or due to intervening events. To make an analogy, suppose one leaves an iron bar outside in the northeastern United States and measures the amount of rust on the bar once per year. One will find that the amount of rust on the bar increases with each passing year. Knowing nothing of chemistry, one might be tempted to conclude that rust is caused by the passage of time per se. In the case of the iron bar, we know this account is incorrect; had the bar been kept in a vacuum, it would not rust at all no matter how long one waits.

In the case of memory, there is little question that many factors affect forgetting above and beyond any effect due to time per se. One could adapt Equation (5.5) to accommodate these factors by allowing ρ to change as a function of variables available at time t. Considering $\mathbf{M}$ as a set of synapses, one might also construct alternative rules for forgetting that allow effects specific to a particular cue and/or a particular response. However, as we will see, there are more fundamental issues with this simple conception of memory as association, so we will not dwell further on this point here.

Forgetting via Stimulus Sampling

Weakening of associations, operationalized as a gradual decrease in the strength of synapses, is not the only way to instantiate forgetting in a simple neural network. Consider Equation (5.6). The term due to weakening of the synapses, ρ^{L-t}, appears with a term relating the similarity of the probe $\mathbf{f}_p$ to each of the cue stimuli in the list $\mathbf{f}_t$. If one provided a probe stimulus that was similar but not identical to one of the cue stimuli in the list, one would expect this to have a measurable effect on memory. For instance, suppose the cue stimulus in an animal conditioning experiment is a pure tone of 440 Hz. One would expect the set of features caused by a similar tone (e.g., 441 Hz) to be greater than the set of features caused by a less similar tone (e.g., 550 Hz). Because this would manifest as changes in the $\mathbf{f}_t^{\mathrm{T}}\mathbf{f}_p$ terms in Equations (5.4) and (5.6), we would expect this to result in more conditioned responding to probes similar to the studied conditioned stimulus. Indeed, it has long been known that one can observe this phenomenon, referred to as stimulus generalization, in animal conditioning experiments (Hull, 1947).

One can use stimulus generalization to construct associative models of forgetting. Stimulus-sampling theory (Estes, 1950, 1955a, 1955b) makes a distinction between the "nominal stimulus" that the experimenter provides and the "functional stimulus" that the research participant experiences. To be more concrete, consider a simple conditioning experiment in which the conditioned stimulus is a 440 Hz tone. The nominal stimulus is the tone itself. A careful experimenter can ensure that the nominal stimulus on each presentation is physically identical. However, no matter how careful the experimenter may be, the functional stimulus experienced by the participant may be meaningfully different from one presentation to the next. For instance, an animal in a Skinner box may have a slightly different posture from one presentation of the nominal stimulus to the next. Or perhaps the animal is more or less attentive to different properties of the nominal stimulus from one presentation to the next. In stimulus-sampling theory the nominal stimulus presented by the experimenter specifies a set of features that *could* be experienced by the participant. On a particular trial, the participant samples from that set of stimulus features to obtain the functional stimulus, which is used to support learning.

It has been said (in a quotation that is often attributed to Heraclitus), that "It is impossible to step into the same river twice." The identity and position of the molecules of water changes continuously from moment to moment. Suppose one steps into a river on two occasions, t_1 and t_2. Although the river at t_1 is not identical to the river at t_2, it is reasonable to say that the similarity of the two rivers, all else equal, is a decreasing function of $t_2 - t_1$. Estes (1955b) proposed that, all else equal, the functional stimulus caused by presentations of the same nominal stimulus at t_1 and t_2 is also a monotonically decreasing function of $t_2 - t_1$. Let us write the functional stimulus caused by nominal stimulus α at time t as $\mathbf{f}_{\alpha,t}$. One can incorporate this assumption into an associative model to enable an account of forgetting without a decrease in the strength of learned associations. Suppose one

learns an association $\mathbf{g}\mathbf{f}^{\mathrm{T}}_{\alpha,t_1}$. Probing with $\mathbf{f}_{\alpha,t_2}$ thus gives $\mathbf{g}$ times a function that decreases with $t_2 - t_1$.

Remarkably, the assumption of gradually changing stimulus features from stimulus-sampling theory from the 1950s has received support from recent neurophysiological studies, at least for some kinds of stimuli. For instance, recent recordings from mouse piriform cortex studied the set of neurons activated by odors during conditioning (Schoonover *et al.*, 2021). The piriform cortex is the first cortical region that receives input from the olfactory bulb, making it roughly analogous to the primary visual cortex for visual images or the primary auditory cortex for auditory stimuli.[2] Because it is so closely related to the sensory receptor itself, it makes sense to think of the activation across the piriform cortex as a direct representation of the sensory stimulus.

The particular recording method that Schoonover *et al.* (2021) used allows for stable recordings of the same neurons over weeks and months. At each stage of the experiment, different odors evoked distinct neural populations. However, the populations that each odor evoked changed continuously over every time period studied. That is, at each time t, one could distinguish $\mathbf{f}_{\alpha,t}$ from $\mathbf{f}_{\beta,t}$. However, $\mathbf{f}^{\mathrm{T}}_{\alpha,t_1}\mathbf{f}_{\alpha,t_2}$ was a decreasing function of $t_2 - t_1$ for all pairs of times considered. Recalling Heraclitus, one might say that the mouse could not smell the same odor twice. This neural phenomenon, referred to as representational drift, is a topic of ongoing research (Mau, Hasselmo, & Cai, 2020; Rule, O'Leary, & Harvey, 2019). Representational drift has been reported, at least under some circumstances, in the visual cortex (Deitch, Rubin, & Ziv, 2021), posterior parietal cortex (Rule *et al.*, 2020), hippocampus (Cai *et al.*, 2016; Mankin *et al.*, 2012; Manns, Howard, & Eichenbaum, 2007; Rubin *et al.*, 2015), and prefrontal cortex (Hyman *et al.*, 2012), as well as piriform cortex.

5.2 Short-Term Memory and Temporal Context Models

The Hebbian associative model from the previous section describes associations between pairs of stimuli. Given a probe stimulus, the model provides a response as output. Although simple and tractable, this model glosses over some fundamental questions about human memory. This section studies models developed largely in response to the free-recall task, which has been an important driver of models of human memory since the 1960s.

In free recall, the participant is presented with a list of stimuli – typically words – one at a time. The participant's task is to recall as many stimuli as possible from the list. In the free-recall task, the participant may recall the words in the order they come to mind (this is in contrast to serial recall where the stimuli must be

2 One may even argue that the piriform cortex is more peripheral than these regions. Information from the retina projects to the visual cortex only after passing through a brain region called the thalamus, which receives information from many sensory modalities. For instance, information from the ear passes through the thalamus on the way to the auditory cortex. In contrast, the piriform cortex is directly connected to the olfactory bulb.

recalled in the order in which they were presented). There are many variants of the free-recall task. In delayed free recall, a distractor task of up to a minute intervenes between the last item in the list and the beginning of the recall period. In the list-before-last paradigm, the participant does not recall the most recent list, but the previous list. In some experiments, participants are given a final free-recall task at the end of the experimental session in which the participant is instructed to recall as many words as possible from all of the preceding lists.

The first problem for the simple Hebbian model that free recall presents is how the task is accomplished at all. The Hebbian model requires a probe to generate a response. What is the probe in free recall? Because the instructions are so general, whatever prompts recall must be internal to the participant. The second challenge for the simple Hebbian model is overwhelming evidence that functional associations are not limited to adjacent items, but are instead distributed very broadly over many items. These findings – reviewed in the next subsection – have led to a very different conception of memory. Rather than a collection of items and associations among them, models originating from the free-recall task have postulated temporally sensitive memory representations that carry information about many items extended over macroscopic periods of time.

5.2.1 The Recency Effect and Two-Store Models

The recency effect refers to the finding that, all else equal, memory is better for information that was presented more recently. In free recall, this manifests as an increase in the tendency to initiate recall at the end of the list (see Figure 5.3 below) as well as higher probability of recall overall. The recency effect can be observed in all of the experimental paradigms that people study with human participants.

The recency effect is especially pronounced in immediate free recall, in which the recall test proceeds just after the last item in the list (Murdock, 1962). In delayed free recall, a delay interval is included during which participants typically perform a distractor task (to prevent them from simply repeating the items in the list to themselves) prior to recalling the words from the list. In delayed free recall the recency effect is sharply attenuated. However, the probability of recall of early items from the list is barely affected relative to immediate free recall (Glanzer & Cunitz, 1966; Postman & Phillips, 1965). In contrast, many other variables (e.g., presenting the words faster or slower, choosing words that are semantically related, having medial temporal lobe amnesia) have a big effect on recall of items from the beginning and middle of the list, but barely any effect on the recency effect (Glanzer, 1972). These observations led researchers to propose that the recency effect draws on a specialized memory store, referred to as short-term store (STS) or short-term memory (Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1980).

The view that memory was divided into distinct stores was hugely influential in the 1970s and 1980s and remains so today. The basic idea (Figure 5.2a) is that STS can store a small number of items with very high accuracy. Items that are

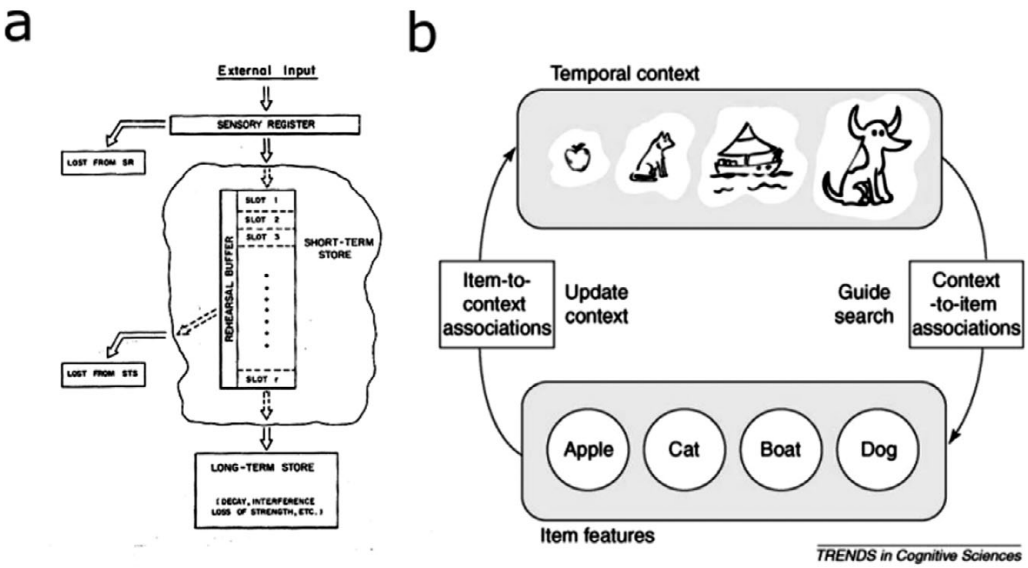


Figure 5.2 *Schematic diagrams for short-term/long-term memory and temporal context models. (a) Models based on a distinction between short-term memory and long-term memory assign different properties to these different stores. The short-term store consists of a rehearsal buffer that contains a small integer number of items with high precision. The long-term store holds a very large number of memory traces with less precision. After Atkinson and Shiffrin (1968). (b) In temporal context models, the currently experienced item activates a set of features on the item layer (bottom). After an item is presented, it activates features that remain active in a gradually changing state of temporal context (top). The context layer cues retrieval via context-to-item associations. The item layer can cause recovery of a previous state of temporal context associated with that item (not shown). After Polyn and Kahana (2008).*

in STS at the time of test are recalled rapidly and with high precision. In addition, a subset of items are passed from STS to a long-term store (LTS). LTS does not have capacity limitations and can store information for a much longer duration. The longer an item spends in STS during study, the greater the probability it is transferred to LTS. A key property of STS is that it is subject to strategic control according to the goals of the participant. For instance, if participants are rewarded based on how many words starting with the letter Q they correctly recall, we might assume that words that start with a different letter are less likely to enter STS and would be forgotten very quickly.

If one specifies a strategy for retaining information in STS it is straightforward to work out (or simulate, if the strategy is very complicated) the probability that an item remains in STS at the time of test. For instance, suppose that each item in a long list enters STS with certainty displacing a random item in STS. If the short-term store can hold N items, where N is much smaller than the

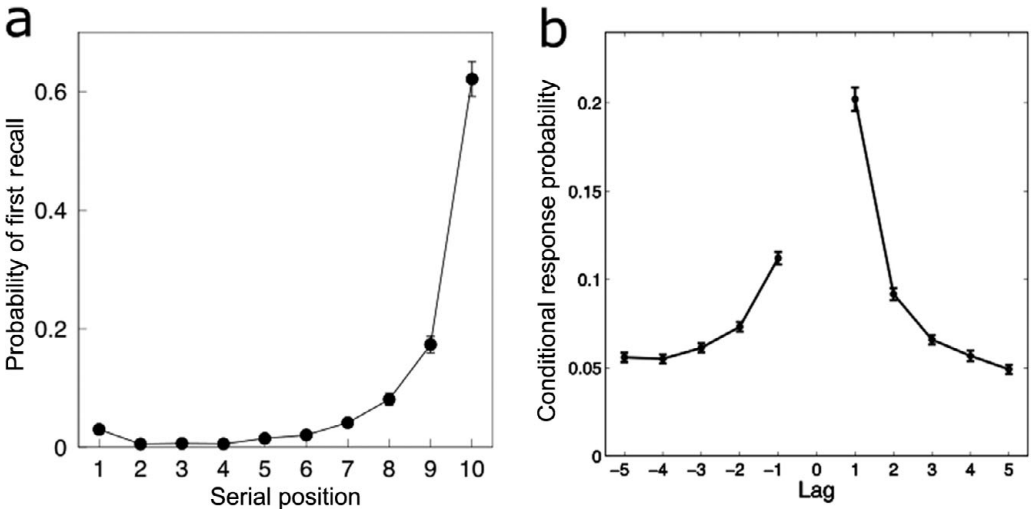


Figure 5.3 *The recency and contiguity effects in free recall. In the free-recall task, participants are presented with a series of stimuli, usually words, and are then asked to recall as many words from the list as possible in the order they come to mind. (a) The recency effect measured by the probability of first recall. The x-axis plots serial position within a list of 10 words. The y-axis gives the probability that the first word the participants said came from each position within the list. In this experiment there is a dramatic recency effect – words from the end of the list are much more likely to be recalled first than words from the beginning or middle of the list. After Howard, Youker, and Venkatadass (2008). (b) The contiguity effect in free recall. Given that a participant has just recalled word i from the list, what is the probability that the next word recalled comes from position* $i + \text{lag}$*? All else equal, participants show a robust tendency to recall words from nearby positions within the list together in recall. The data in this figure is averaged over many experiments. After Kahana (2012).*

number of items in the list, then the probability that an item already in STS is replaced by an incoming item is $1/N$. The probability that the item already in STS persists in STS after a new item enters STS is thus $1 - 1/N$. At the end of a list of L items, the probability that the ith item is still in STS at the time of test is $(1 - 1/N)^{L-i}$, leading to a recency effect. Note that although this function decays exponentially, recency due to STS has different properties than recency due to exponential weight decay [Equation (5.6)]. First, the quantity that is decaying is a probability rather than a strength per se. This probability gives the proportion of trials where the item is available for recall from STS; on trials where the item is not available, there is zero probability of retrieval from STS. This is distinct from a situation where the weights give a small but reliable signal. Second, although the probability of any one item remaining in STS may be a decreasing function, it should be kept in mind that the number of items in STS depends only on its capacity N (assuming the list has more than N items).

One can similarly work out probabilities for the amount of time a word spends in STS (recall that the probability of transfer to LTS goes up with time spent in STS). Coupled with a specfication of LTS one can make predictions for many observable properties of memory retrieval, resulting in a very detailed description of immediate and delayed free recall, including but not limited to the recency effect.

A major challenge to the two-store account of recency came from a modification to the free recall paradigm referred to as continual distractor free recall (CDFR). Recall that in immediate free recall the recall test follows shortly after the last item in the list. According to STS-based accounts, the recency effect in immediate free recall happens because the items from the end of the list are still available in STS. In delayed free recall, a distractor task follows the last item on the list before the recall test. The recency effect is attenuated in delayed free recall. According to STS-based accounts, this is a consequence of the distractor task pushing list items out of STS. In CDFR, a distractor task follows each item in the list, not only the last item. Perhaps surprisingly, there is a pronounced recency effect in continual distractor free recall relative to delayed free recall (Bjork & Whitten, 1974; Glenberg *et al.*, 1980). This finding was not predicted by the STS-based account of recency and is difficult to reconcile with an account of recency solely based on STS (Davelaar *et al.*, 2005; Lehman & Malmberg, 2012).

5.2.2 The Contiguity Effect Across Delays

As a thought experiment, try the following memory experiment on yourself. Answer the following question: WHAT DID YOU MOST RECENTLY HAVE FOR BREAKFAST?[3] Most people, when answering this question, do not merely generate a verbal response (e.g., "toast") but experience a vivid recollection of the event in the process of answering the question. For instance, while writing this (in the afternoon), in answering the question about breakfast, I spontaneously remembered where I sat down (kitchen table with the window to my right), the hopeful look on my dog's face, and the news I read on my phone. I can take another moment and search my memory to vividly remember events that happened shortly before eating breakfast (putting the coffee on the stove, putting bread in the toaster) and shortly after (finishing my coffee in the backyard with my dog).

The "kind of memory" that supports vivid recollection of events from one's life is referred to as episodic memory (Tulving, 1983). Episodic memory has been extensively studied over the last several decades. For the present purposes we note that episodic memory is believed to be closely related to a phenomenon referred to as the contiguity effect. In free recall, the contiguity effect (Figure 5.3b) manifests as the finding that (all else equal) if a participant has just recalled a word from the list, the next word that participant recalls tends to come from a nearby position in

3 If you are eating breakfast while reading this you can substitute the question WHAT DID YOU MOST RECENTLY HAVE FOR DINNER?

the list (Kahana, 1996). In memory experiments with a probe (e.g., cued recall), the contiguity effect manifests as the finding that the probe tends to bring to mind other items that were close together in time. For instance, in cued recall, when a participant recalls a word that was not the correct response to the probe, that erroneous word tends to come from a pair that was presented nearby in the list. The contiguity effect is not limited to experiments with words as stimuli and is indeed quite general (Healey, Long, & Kahana, 2018).

Note that the episodic memory for today's breakfast illustrates the contiguity effect. Sitting down at the table, giving my dog a piece of sausage and reading about terrible events unfolding overseas were not actually simultaneous but were relatively close together in time (probably tens of seconds). The other events I retrieved – putting the bread in the toaster and finishing the coffee in the backyard – were each separated by several minutes from breakfast per se. Consistent with this intuition, the contiguity effect is observed in the laboratory in CDFR experiments where the items are separated by tens of seconds. The contiguity effect can also be observed over much longer time scales – hundreds of seconds in final free recall (Howard, Youker, & Venkatadass, 2008), hours in experiments using mobile phones to administer a list as participants went through their daily lives (Mack *et al.*, 2017) and even much longer periods of time in retrieving news events (Uitvlugt & Healey, 2019).

One may think of the contiguity effect as analogous to the recency effect, but taken from a different temporal reference frame. The recency effect describes the availability of items in memory as a function of their temporal proximity to the present. In contrast, the contiguity effect describes the availability of items in memory as a function of their temporal proximity to a remembered moment from the past. This analogy between recency and contiguity suggested a different class of models for memory, which we turn to in the next subsection.

5.2.3 Temporal Context Models

In this subsection we describe the memory representations of a class of models referred to as temporal context models (TCMs, Howard & Kahana, 2002; Sederberg, Howard, & Kahana, 2008; Polyn, Norman, & Kahana, 2009). These models were originally developed to account for recency and contiguity effects in free recall. TCMs have since been applied to other episodic memory tasks, and even memory tasks that are not considered to tap episodic memory (Logan, 2021). In this subsection we will describe the basic properties of these models and how they result in properties of memory. We will discuss neuroscientific work inspired by TCMs before describing some fundamental limitations that follow from the form of temporal context.

Temporal context models make three important conceptual changes relative to the models we have considered thus far in this chapter. First, these models hypothesize a vector representation of temporal context that changes gradually from moment to moment. We will specify this in more detail below. For now,

we note that the temporal context vector shares at least some features with the content of short-term store. Second, temporal context models do not attribute behavioral associations between items – such as the contiguity effect – to direct connections formed between item representations [as in Equation (5.1)]. Rather, functional associations in temporal context models are mediated by items' effects on temporal context and a temporal context's ability to cue retrieval of items. Third, temporal context models assume that it is possible to reinstate a previous state of temporal context. This "jump back in time" is hypothesized to be associated with the experience of episodic memory.

Two Interacting Vector Spaces: Items and Contexts

In TCMs, there are two interconnected vector spaces (Figure 5.2b). One vector space, which we will sometimes refer to as the item space, is activated by items that are currently available, either by virtue of having been presented by the experimenter or by virtue of having been recalled by the participant. We refer to the cognitive representation of specific items as vectors $\mathbf{f}$ and the vector corresponding to the item presented at time step t as $\mathbf{f}_t$. The other vector space, which we will sometimes refer to as the context space, maintains a state of temporal context. We will refer to the state of temporal context at time t as $\mathbf{c}_t$. Temporal context is affected by items; the input at time t, $\mathbf{c}_t^{\mathrm{IN}}$, is caused by $\mathbf{f}_t$, the item available at time t.

Temporal context evolves gradually, retaining information contributed by recent items:

$$\mathbf{c}_t = \rho \mathbf{c}_{t-1} + \mathbf{c}_t^{\mathrm{IN}}. \tag{5.7}$$

That is, at each time step t, the new state of temporal context is given by ρ times the previous state of temporal context, plus the input caused by $\mathbf{f}_t$, $\mathbf{c}_t^{\mathrm{IN}}$. As before, $0 < \rho < 1$ so that in some formulations, ρ is allowed to vary as a function of time (for instance to normalize the context vector) and/or can vary for different components of the context vector as attention to different features changes (e.g., due to different encoding tasks). We assume that on the initial presentation of an item in a randomly assembled list of words, the inputs caused by each item $\mathbf{c}^{\mathrm{IN}}$ are uncorrelated with one another and treat them as random vectors. Equation (5.7) shows that information caused by a particular item persists after it is presented. Recursively unwinding Equation (5.7), we find

$$\mathbf{c}_t = \sum_{\tau=0}^{\infty} \rho^{t-\tau} \mathbf{c}_{t-\tau}^{\mathrm{IN}}. \tag{5.8}$$

That is, the input pattern $\mathbf{c}^{\mathrm{IN}}$ caused by an item decays exponentially as additional items are presented.

At any particular moment, recall is cued by the current state of temporal context via an associative matrix $\mathbf{M}^{CF}$ that connects the context layer (containing context vectors $\mathbf{c}$) to the item layer (containing item vectors $\mathbf{f}$). Analogous to our simple Hebbian model [Equation (5.1)], the basic formulation provides an outer product

association between the context available prior to presentation of the current item and the item itself:

$$\Delta \mathbf{M}^{CF} = \mathbf{f}_t \mathbf{c}_{t-1}^{\mathrm{T}}. \tag{5.9}$$

This shift in indices ensures that the temporal context that cues $\mathbf{f}_t$ does not include information $\mathbf{c}_t^{\mathrm{IN}}$ that the item itself caused.

Equation (5.9) resembles Equation (5.1) in that it associates two patterns via an outer product. However, rather than associating two items $\mathbf{f}$ and $\mathbf{g}$, $\mathbf{M}^{CF}$ associates a context vector to an item vector. The context-to-item association means that a probe context activates each item in the list to the extent the probe context resembles that item's encoding context. By analogy to Equation (5.4):

$$\mathbf{M}^{CF}\mathbf{c}_p = \sum_t \left(\mathbf{c}_{t-1}^{\mathrm{T}}\mathbf{c}_p\right)\mathbf{f}_t. \tag{5.10}$$

Because context changes gradually, this typically results in a weighted sum of many items. Temporal context models use a retrieval rule to probabilistically select an item for recall. These mechanisms are sometimes quite elaborate; the key feature they share is that the probability of recalling a particular item at a particular retrieval attempt depends not only on the degree to which it is activated, but also on the activation of the other items in the list. That is to say, items compete to be retrieved.

Recency Effect

We are in a position at this stage to understand why TCMs predict recency effects in immediate and delayed free recall. Combining Equations (5.8) and (5.10) we find, under the assumption that the $\mathbf{c}^{\mathrm{IN}}$ during initial study of a random list are orthogonal to one another, that probing with the context available at the end of the list, $\mathbf{c}_L$, gives back the items from the list weighted exponentially:

$$\mathbf{M}^{CF}\mathbf{c}_L \propto \sum_t \rho^{L-t+1}\mathbf{f}_t. \tag{5.11}$$

The exponential decay clearly provides a large advantage to items from the end of the list, leading naturally to a robust recency effect. Introducing a delay D takes $\mathbf{c}_L \rightarrow \rho^D \mathbf{c}_L +$ distractors, where the distractors ought to be orthogonal to the list items. This reduces the difference in activation between the last items in the list and earlier items, resulting in a decrease in the magnitude of the recency effect.

5.2.4 Contiguity Effect

Thus far we have considered only the case where the input patterns $\mathbf{c}^{\mathrm{IN}}$ caused by the items in the list are orthogonal to one another. In this subsection we study the effects of relaxing this assumption. To make the ideas clear, let's repeat an item at the end of a very long list of unrepeated items and see how the resulting context cues the neighbors of the repeated item. We consider two possibilities. In the first case, the repeated item simply causes the same input that it did during the initial

presentation of the list. In the second case, we consider the case that the repeated item recovers the temporal context available when it was intially presented; that the repeated item causes a jump back in time. We will find that these two hypotheses result in very different qualitative properties.

Let us label the time index at which an item is repeated as r, the position at which the repeated item was initially presented as i, and study the ability of $\mathbf{c}_r^{\text{IN}}$ to cue items near i, $\mathbf{f}_{i+\text{lag}}$. We assume that r is far in the future so that we can neglect $\mathbf{c}_{i+\text{lag}}^{\text{T}}\mathbf{c}_{r-1}$ and restrict our attention to $\mathbf{f}_{i+\text{lag}}^{\text{T}}\mathbf{M}^{CF}\mathbf{c}_r^{\text{IN}}$. Suppose that the repeated item simply causes the same input at time step r that it did when it was initially presented at time step i. Because $\mathbf{c}_i^{\text{IN}}$ persisted after time step i [see Equations (5.7) and (5.8)], this results in similarity to the context states that followed time step i. This similarity decreases exponentially with lag > 0. Put another way, because temporal context contains information from recently presented items, $\mathbf{c}_i^{\text{IN}}$ is similar to the temporal context of items for which i was in the recent past. However, the same is not true for items that *preceded* item i. For lag ≤ 0, information retrieved by item i is not in the recent past – item i has not been presented yet and there is no way the participant should be able to predict a word in a random list. Putting these considerations together, we find that if $\mathbf{c}_r^{\text{IN}} = \mathbf{c}_i^{\text{IN}}$:

$$\mathbf{f}_{i+\text{lag}}^{\text{T}}\mathbf{M}^{CF}\mathbf{c}_i^{\text{IN}} = \begin{cases} 0, & \text{lag} \leq 0 \\ \rho^{\text{lag}}, & \text{lag} > 0 \end{cases}. \tag{5.12}$$

That is, if at time step r, the item at time step i simply recovers the same input it caused during encoding, $\mathbf{c}_r^{\text{IN}} = \mathbf{c}_i^{\text{IN}}$, this results in an asymmetric functional association to its neighbors.

Now let's consider the case in which the repeated item recovers the state of context available when it was initially presented, $\mathbf{c}_r^{\text{IN}} = \mathbf{c}_{i-1}$. This context includes information caused by the items that preceded item i. This information also persists in the temporal context after item i was presented. Noting that the inner product is symmetric, $\mathbf{v}^{\text{T}}\mathbf{u} = \mathbf{u}^{\text{T}}\mathbf{v}$, we conclude that in this case

$$\mathbf{f}_{i+\text{lag}}^{\text{T}}\mathbf{M}^{CF}\mathbf{c}_i \propto \rho^{|\text{lag}|}. \tag{5.13}$$

That is to say, retrieving the previous state of temporal context results in a symmetric association that falls off exponentially as a function of |lag|.

In most free-recall experiments, the shape of the contiguity effect includes a contiguity effect in both the backward and forward direction, with a reliable advantage for forward transitions (Figure 5.3b is representative). In TCMs, the pattern retrieved by item i when it is re-experienced at time step r is a mixture of these two patterns:

$$\mathbf{c}_r^{\text{IN}} = (1-\gamma)\,\mathbf{c}_i^{\text{IN}} + \gamma\mathbf{c}_i. \tag{5.14}$$

The value of γ can be estimated from the data and is believed to vary not only from participant to participant but also from one retrieval to the next. This makes sense of the finding that episodic memory retrieval – presumably related to the recovery

of a previous state of temporal context – does not always succeed. This property of episodic memory is familiar to anyone who has bumped into a familiar person in a public place (e.g., a grocery store) . . . but been unable to actually remember any details of the person's identity.

5.2.5 Neural Evidence for Temporal Context Models

Temporal context models have benefited from a relatively close connection to work in cognitive neuroscience. After all, if the long-term goal of this kind of modeling is to develop a more-or-less literal model of the computations that take place in the brain during memory encoding and retrieval, it is essential to compare hypotheses to the activity of neurons in the brain. We briefly point to three pieces of evidence that speak to the utility of TCMs in making sense of human and also animal neuroscience.

First, the division of $\mathbf{c}^{\text{IN}}$ into two components with distinct properties [see Equation (5.14)] has been very productive in explaining otherwise isolated findings in neuropsychology and cognitive neuroimaging. To take a simple example, imagine if it were possible to alter γ across experimental groups. A group with a lower value of γ ought to have difficulties with vivid episodic memory recall, but also show a more asymmetric contiguity effect in free recall. This finding has been observed with patients with medial temporal lobe amnesia (Palombo *et al.*, 2019), electrical stimulation to the entorhinal cortex (Goyal *et al.*, 2018), and participants who are experiencing cognitive declines with aging, perhaps leading to Alzheimer's disease (Quenon *et al.*, 2015; Talamonti *et al.*, 2021). Moreover, according to the models, retrieved temporal context ought to be preferentially involved in particular sorts of memory. Consider an experiment where participants learn pairs separated by long periods of time, ABSENCE HOLLOW . . . HOLLOW PUPIL. If the second presentation of HOLLOW can cause recovery of its previous context (i.e., the $\mathbf{c}^{\text{IN}}$ caused by ABSENCE), then ABSENCE in effect becomes part of the temporal context for PUPIL. If $\gamma = 0$, the model can still learn the pairwise associations using the forward part of the contiguity effect. Indeed, normal human participants generalize ABSENCE PUPIL associations even though ABSENCE and PUPIL were never experienced nearby in time. As it turns out, lesions to a brain region called the hippocampus – which is believed to be important in episodic memory – cause a deficit in these bridging or "transitive" associations in rodents while leaving the pairwise associations unaffected (Bunsey & Eichenbaum, 1996), just as if the hippocampus is responsible for causing a recovery of temporal context. A number of neuroimaging studies have looked at similar experimental paradigms in humans, showing that the hippocampus and hippocampal–prefrontal interactions are important in these transitive associations (Zeithamova, Dominick, & Preston, 2012).

One can also measure direct neural predictions from TCMs. The most characteristic prediction is the existence of a temporal context vector $\mathbf{c}$, which should show temporal autocorrelation extending over macroscopic periods of time – at

least tens of seconds. One can construct a vector of brain activity using many different methods. For instance, it is practical to record simultaneously from many individual neurons at once. Taking the number of spikes for each of N neurons averaged over, say, a one second interval gives an N-dimensional vector. One can then compute a temporal autocorrelation function by comparing response vectors from neighboring time points. This type of analysis has shown robust evidence for signals that are autocorrelated over seconds, minutes, and even hours or days in a number of brain regions, notably the hippocampus and prefrontal cortex (Cai *et al.*, 2016; Hyman *et al.*, 2012; Mankin *et al.*, 2012). These studies have focused on rodents because of the array of systems neuroscience tools that can be brought to bear in rodents, but analogous results have been found with human fMRI (Hsieh *et al.*, 2014).

The most characteristic prediction of TCMs is that the state of temporal context should be recovered when an episodic memory is retrieved [see Equation (5.13)]. When item i is repeated at some later time step r, and causes an episodic memory, the context at time step r should resemble the context *prior* to the context at time step i. This is nontrivial; any neural information that was caused by item i during study can only be observed after its original presentation. There is evidence from invasive human recordings of this phenomenon in several human memory paradigms (Folkerts, Rutishauser, & Howard, 2018; Manning *et al.*, 2011; Yaffe *et al.*, 2014), fMRI studies of free recall (Chan *et al.*, 2017), and real-world memory extended over hours and days and weeks (Nielson *et al.*, 2015).

5.2.6 Memory is Scale-Invariant; Exponential Functions are Not

In our discussion of models of short-term memory, we noted that the failure of short-term memory models to account for the long-term recency effect and long-term contiguity effects was a serious problem for those models. It is true that TCMs are better able to account for those phenomena. In STS-based models, the probability that an item is perfectly represented in STS falls off exponentially. As time passes, STS provides zero information about the item on an increasingly high proportion of trials. In contrast, in TCMs the information about an item falls off exponentially with time, but is reliable across trials. With a bit of resourcefulness and a few free parameters, one can exploit this property to provide a reasonable fit to experimental data from continuous distractor free recall. But this account is still theoretically unsatisfactory, as we shall see shortly.

As discussed above, a great deal of evidence suggests that recency and contiguity effects not only persist across a delay interval in CDFR, but are observable at an extremely wide range of time scales (Figure 5.4c provides a particularly striking example). This suggests that the memory representations governing recency and contiguity effects are scale-invariant (Chater & Brown, 2008). A function is said to be scale-invariant if it is unaffected by rescaling the input up to a scaling factor. That is, a function $y(x)$ is said to be scale-invariant if stretching or compressing its input by a constant, $x \rightarrow ax$, results in the same function up to a constant

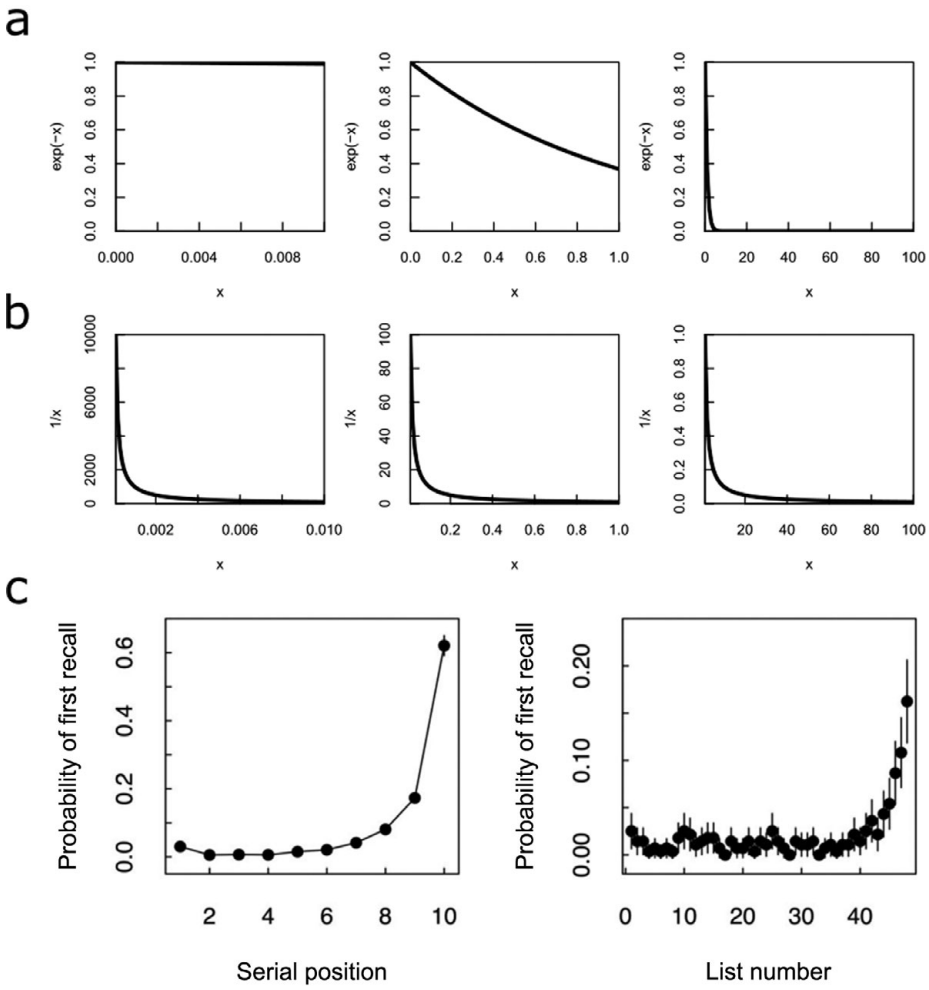


Figure 5.4 *Scale-invariant memory. Consider taking a variable x and rescaling it* $x \rightarrow ax$. *(a) An exponential function* e^{-x} *zoomed in over different ranges of x. (b) A power-law function* x^{-1} *zoomed in over different ranges of x. Starting from the middle panel, where x is shown over the range 0 to 1, the left panels show the functions rescaled by zooming in on x by a factor of 100; the right panels show the functions zoomed out by a factor of 100. Note that the exponential function has very different properties across scales. In contrast, the power-law function has the same shape up to a scaling factor (note the change in the y-axis) regardless of the scale over which it is examined. (c) The recency effect in human memory persists across time scales. Left: Memory tested on the scale of seconds. Right: Memory tested on the scale of minutes. Participants studied lists of words. The left panel shows the probability that the first word that came to mind in a free-recall task came from each position within the list. After learning 48 lists, participants were asked to recall all the words they could remember from all the lists in the experimental session. The right panel plots the probability that the first word they recalled came from each* list *in the session. Note that the function has a similar shape across very different time scales. After Howard, Youker, and Venkatadass (2008).*

term that depends only on a: $y(ax) = f(a)y(x)$. This property is true of power-law functions that govern, say, electrical potential as a function of distance from a charged particle, or the gravitational field as a function of distance from a massive object in Newtonian gravity. We can easily convince ourselves of this property by noting that if $y(x) = x^{-1}$, then $y(ax) = a^{-1}y(x)$, satisfying the constraint. Figure 5.4b illustrates this property for $y(x) = x^{-1}$ by rescaling the x-axis.

The exponential functions generated by TCMs are decidedly not scale-invariant. Note that $\rho^x = e^{-x}$ if we choose $\rho = 1/e$. More generally, $\rho^x = e^{-sx}$ if $\rho = e^{-s}$ so that $s = -\log \rho$. Thus, choosing a ρ is equivalent to specifying a rate constant s (or a time constant $1/s$) for an exponentially decaying function. Figure 5.4a shows the function $y(x) = e^{-x}$ rescaled over the same range of values as the power-law function. When x is much less than one (left panel), the exponential function appears linear. This follows from the Taylor series expansion of the exponential function:

$$e^{-\Delta} = 1 - \Delta + \cdots, \tag{5.15}$$

where additional terms include higher powers of Δ multiplied by e^{-x}. As we zoom out (right panel), the exponential function comes to approximate a delta function centered at zero. Note that in both of these two regimes $x \ll 1$ and $x \gg 1$, the exponential function is useless for expressing a recency effect. Mapping x to recency, when x is small, there is no forgetting because all points are associated with a high nearly constant value. When x is large, almost all points (excluding zero) are mapped to a low nearly constant value.

This rescaling is not an academic exercise. CDFR approximates rescaling of experience. Insertion of a delay of duration D between each item and at the end of the list approximates taking $\rho \rightarrow \rho^D$, so that the relative delay between serial positions relative to the time of retrieval becomes effectively larger. From this it is clear that, although one may be able to approximate experimental data in restricted cases, the machinery of the temporal context vector specified by Equation (5.7) will eventually break down.

5.3 Scale-Invariant Temporal History

Thus far, we have considered models based on more or less complicated implementations of the idea of association. In the case of the Hebbian association model, the association is distributed across the entries in a matrix corresponding roughly to the set of synapses between items. In temporal context models, associations between items are mediated by temporal context, a representation of the recent past in which previous events decay gradually. These models share an implicit assumption that the goal of memory is to express relationships as a scalar value. That is, we can talk about the relationship between, say, ABSENCE and HOLLOW only in terms of the magnitude of the connection between them. Given two pairs, ABSENCE–HOLLOW and PUPIL–RIVER, the simple Hebbian model

does not have any mechanism to convey information about whether one pair was learned before or after the second pair. Yes, one might note that the ABSENCE–HOLLOW association is stronger than the PUPIL–RIVER association and use this to infer that ABSENCE–HOLLOW was more recent, but this inference would break down if, for instance, the participant was paying less attention when PUPIL–RIVER was presented, or if ABSENCE–HOLLOW was presented multiple times.

Similar arguments apply to TCMs. Although temporal relationships can be inferred indirectly from the magnitude of the associations between multiple words, there is no explicit information about the direction of time contained in $\mathbf{c}_t$. Consider two context vectors $\mathbf{c}_t$ and $\mathbf{c}_{t+\text{lag}}$. The direction of the difference between these two vectors, $\mathbf{c}_{t+\text{lag}} - \mathbf{c}_t$, depends on the particular choice of items presented during the interval specified by lag rather than the time per se. Moreover, as with simple Hebbian models, repeated items can make even the magnitude of these vectors ambiguous. The goal of the representation used in this section is to build a replacement for the temporal context vector. We desire that this representation carries explicit information about temporal relationships. We also desire that this representation can be used to build scale-invariant models of memory.

Understanding vectors as activated populations of neurons, the simple Hebbian model and temporal context vectors distribute "what" information about the stimuli that are experienced across populations of neurons. Different basis vectors of the space correspond to different properties of stimuli. The temporal context vector provides decaying "what" information "smeared" over the recent past. The strategy of this approach is to construct a population of neurons that not only represent information about what has happened in the recent past, but to distribute information about when it happened across different neurons. That is, our computational goal is to estimate the recent past as a function of time. Figure 5.5 provides an illustration and introduces notation. In this section we describe a specific solution to this problem that has found considerable empirical support from data from both psychology and neuroscience.

Let us suppose that the world provides a continuous stream of input $f(t)$. Like the set of vectors corresponding to a list of words, f is in general vector-valued but we will suppress vector notation for now. Consider the problem of an observer having examined f up to a particular point t. We will refer to the history leading up to this moment t as $f_t(\tau)$, where τ runs from zero to ∞ and $\tau = 0$ corresponds to the present. Our goal is to construct an estimate of the history leading up to time t as $\tilde{f}_t(\overset{*}{\tau})$. We desire that this estimate approximates reality – with an error that is comparable across time scales – and is also a computation that could be implemented by neural circuits. The next subsection introduces a specific method that has these properties (proposed by Shankar & Howard, 2012). Subsequent subsections demonstrate that it is straightforward to build not only temporal context models out of this form of representation, but also other more "cognitive" models as well. Finally, we touch on a wealth of neuroscience work that suggests populations of neurons like those proposed for $\tilde{f}_t(\overset{*}{\tau})$.

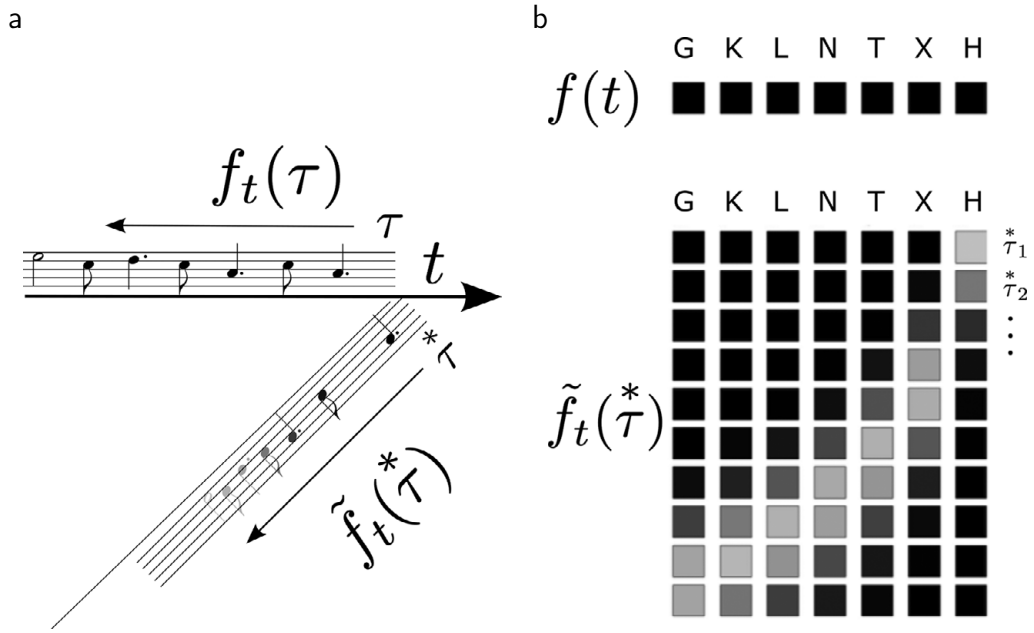


Figure 5.5 *Scale-invariant temporal history. (a) Cartoon illustrating the goal of the scale-invariant temporal history. At time t, the history leading up to the present is given by $f_t(\tau)$. The argument τ runs from 0 to ∞. The goal of the representation of temporal history is to construct at each moment a record of the recent past as a scale-invariant temporal history. This history is compressed in that it has less temporal resolution for events further in the past. (b) Schematic of the temporal history at a single moment shortly following presentation of a list* G K L N T X H. *Each box gives the activation of a "unit" at time t. Lighter boxes indicate higher activation. Black boxes indicate zero activation. Top: As in TCMs, the input pattern $f(t)$ is a vector over items. Here we assume that each item has an orthogonal representation; the features are sorted on their order of past presentation for ease of visualization. Because we take t to be shortly after presentation of the last item in the list, there is no activation in $f(t)$. Bottom: The scale-invariant representation retains information about the past leading up to the present. Here "columns" are organized so that they correspond to the same features as in $f(t)$. Columns correspond to "what" information. Rows correspond to "when" information. For instance, in the top row, only the column corresponding to* H, *the last item in the list, is active. For rows representing information further in the past, several items are active (note that the peaks for* K *and* L *overlap). The curvature in the peak of activation across the list items is a consequence of the logarithmic compression of the internal time axis. The grayscale changes across rows for ease of visualization. In actuality, the peak of a stimulus a time τ in the past goes down like τ^{-1}.*

5.3.1 Estimating Temporal Relationships Using the Laplace Transform

This section describes a method for estimating $\tilde{f}_t(\overset{*}{\tau})$ based on Laplace transforms that was proposed by Shankar and Howard (2012). First let us write a continuous version of Equation (5.7). For reasons that will become clear, we change notation such that the temporal context vector $\mathbf{c}_t$ is written as $F(t)$ and the input to the context vector $\mathbf{c}_t^{\text{IN}}$ is written as $f(t)$. We take both of these to be vector-valued but will suppress the vector notation for present. Defining $s = -\log \rho$, this is just a continuous version of Equation (5.7):

$$\frac{dF}{dt} = -sF + f(t). \tag{5.16}$$

Solving Equation (5.16) we find, in the general case:

$$F_t(s) = \int_0^\infty e^{-s\tau} f_t(\tau) d\tau. \tag{5.17}$$

Comparing this to Equation (5.8) we see a close correspondence between $\mathbf{c}_t$ and F_t if we make the identification $\rho = e^{-s}$. In contrast to the TCMs we discussed in Section 5.2.3, we do not understand s as a parameter to be estimated from the data of a particular experiment, but as a continuous variable. To be concrete, we can imagine that we have an ensemble of units, each with a different value of s.

Continuous s Enables Information About Continuous Time

Treating s as a continuous variable allows us to reconstruct information about the value of $f_t(\tau)$ at different values of τ. With any particular value s_1, $F_t(s_1)$ captures information about the past history $f_t(\tau)$ up to a time scale on the order of $\tau_1 = 1/s_1$. If we chose a different value s_2, $F_t(s_2)$ would capture information up to $\tau_2 = 1/s_2$. For simplicity, let's assume that $\tau_1 < \tau_2$. Consider the properties of the exponential function illustrated in Figure 5.4. For values of τ much less than τ_1, both $F_t(s_1)$ and $F_t(s_2)$ weight $f_t(\tau)$ by similar amounts. Similarly, for values of τ much greater than τ_2, both of the exponential functions have decayed to zero and neither $F_t(s_1)$ nor $F_t(s_2)$ carries information about $f(\tau)$ in that interval. However, consider how the two values of F vary as τ increases from τ_1 to τ_2 (recall that $\tau_1 < \tau_2$). As τ passes through τ_1, the contribution of $f_t(\tau)$ to $F_t(s_1)$ rapidly decreases. However, the exponential for $F_t(s_2)$ decays less steeply in this region, so that the contribution of these values to $F_t(s_2)$ is greater. We conclude that one can infer something about the values of $f_t(\tau)$ in a region specified by τ_1 and τ_2 by observing the difference between $F_t(s_1)$ and $F_t(s_2)$. Given many values of s we can infer $f_t(\tau)$ at many values of τ.

More formally, we can note that $F_t(s)$ from Equation (5.17) describes the real Laplace transform of $f_t(\tau)$. The Laplace transform is invertible; if we know the value of $F_t(s)$ precisely with every real value of s from 0 to ∞, then we can specify

$f_t(\tau)$ precisely for every value of τ from 0 to ∞. We will restrict our attention to real positive values of s.[4]

Approximately Inverting the Laplace Transform

Now that we've established that $F_t(s)$ carries information about the time of past events $f_t(\tau)$, we need to determine how to extract that information. Knowing that $F_t(s)$ is the real Laplace transform of $f_t(\tau)$ suggests a strategy – simply invert the Laplace transform. That is, $F_t(s)$ provides a memory for the past leading up to the present $f_t(\tau)$. After inverting the Laplace transform, we would obtain an estimate of the actual history, which we write as $\tilde{f}_t(\overset{*}{\tau})$. Over the years, many methods for the inverse Laplace transform have been proposed. We focus on the Post approximation (Post, 1930), which is relatively straightforward to implement in neural circuits and has some computational properties that are advantageous in describing psychological and neurophysiological results.

To approximately invert the transform, we define a mapping $\overset{*}{\tau} \equiv k/s$, where k is an integer to be approximated from the data. At each moment, the value of $\tilde{f}$ at each value of $\overset{*}{\tau}$ is computed as

$$\tilde{f}_t(\overset{*}{\tau}) \equiv \mathbf{L}_k^{-1} F_t(s) = C_k s^{k+1} \frac{d^k}{ds^k} F_t(s). \tag{5.18}$$

The derivative on the right-hand side is to be taken in the neighborhood of the value of $s = k/\overset{*}{\tau}$. C_k is a constant that ensures that the sign and magnitude of $\tilde{f}_t(\overset{*}{\tau})$ corresponds to the sign and magnitude of $f_t(\tau)$. The operator $\mathbf{L}_k^{-1}$ includes a computation of the kth derivative with respect to s.[5] In the limit as $k \to \infty$, the Post approximation becomes the inverse transform and $\tilde{f}_t(\overset{*}{\tau} = \tau) = f_t(\tau)$. However, for finite k, there is a temporal blur introduced. $\tilde{f}_t(\overset{*}{\tau})$ is equal to an average of $f_t(\tau)$ in the neighborhood around $\tau = \overset{*}{\tau}$. Suppose $f_t(\tau)$ is a delta function at a particular time τ_o in the past. Then

$$\tilde{f}_t(\overset{*}{\tau}) = C_k \, s^{k+1} \, \frac{d^k}{ds^k} e^{-s\tau_o} \tag{5.19}$$

$$= C_k \, s^{k+1} \tau_o^k \, e^{-s\tau_o} \tag{5.20}$$

$$= C_k \, \frac{1}{\overset{*}{\tau}} \left(\frac{\tau_o}{\overset{*}{\tau}} \right)^k e^{-k\left(\frac{\tau_o}{\overset{*}{\tau}}\right)}. \tag{5.21}$$

The constant C_k includes a factor of -1^k so that the right-hand side of this expression is positive for all k. The function on the right-hand side of Equation (5.21) is a product of a growing power law and a decreasing exponential, resulting

4 Negative real values of s would be neurally unreasonable. We ignore complex s for simplicity.

5 Given a discrete set of s values, $\mathbf{L}_k^{-1}$ can be understood as a matrix L_{ij} that maps $F(s_j)$ onto $\tilde{f}(\overset{*}{\tau}_i)$, with a matrix implementation of the discrete derivative.

in a function that has a single peak. Freezing time at a particular τ_o and looking across all $\overset{*}{\tau}$, the peak comes at $\overset{*}{\tau} = \tau_o \frac{k}{k+1}$. Fixing a particular $\overset{*}{\tau}$ and observing it through time as τ_o changes, the peak comes at $\tau_o = \overset{*}{\tau}$. The most important property of this expression is that the right-hand side depends on the time τ_o only through ratio $\tau_o/\overset{*}{\tau}$. Because of the linearity of Equation (5.17) and the linearity of $\mathbf{L}_k^{-1}$, we can write an expression for any history $f_t(\tau)$ as

$$\tilde{f}_t(\overset{*}{\tau}) = \int_0^\infty C_k \frac{1}{\overset{*}{\tau}} \left(\frac{\tau}{\overset{*}{\tau}}\right)^k e^{-k\frac{\tau}{\overset{*}{\tau}}} f_t(\tau)\, d\tau \tag{5.22}$$

$$= \int_0^\infty \frac{1}{\overset{*}{\tau}} \Phi_k\left(\frac{\tau}{\overset{*}{\tau}}\right) f_t(\tau)\, d\tau \tag{5.23}$$

$$= \int_0^\infty \Phi_k(x)\, f_t\left(\overset{*}{\tau} x\right)\, dx, \tag{5.24}$$

where we have defined $\Phi_k(x) \equiv x^k e^{-kx}$ and changed variables to $x \equiv \frac{\tau}{\overset{*}{\tau}}$ in the last line.

A Note on Biological Realism

As we will see later, these equations provide a reasonable description not only of a memory representation that can be used to describe behavior in a range of memory tasks, but also of neurophysiological data from a number of brain regions. The equations are in principle computable by neurons – Equation (5.16) simply requires slow time constants and it has long been known that the brain can compute derivatives needed to implement $\mathbf{L}_k^{-1}$. How literally should one take these equations? There is certainly a level of precision at which these equations are not a correct description of the firing rate of neurons. The author of this chapter encourages the reader to take these equations seriously, but not literally.

For instance, Equation (5.16) describes an instantaneous reaction to an input in continuous time. If one understands $f(t)$ as a stimulus under external control, this cannot be literally true. Moreover, there are a number of ways in which the brain could implement the slow rate constants in Equation (5.16), including recurrent connections, metabotropic glutamate receptors (Guo *et al.*, 2021), and feedback loops between spiking and intrinsic currents (Egorov *et al.*, 2002; Tiganj, Hasselmo, & Howard, 2015). These mechanisms would all have slightly different properties that would deviate from Equation (5.16). However, the larger point that firing for a population of neurons decays roughly exponentially following a triggering stimulus with a broad range of time constants may still be true.

Similarly, the inverse operator $\mathbf{L}_k^{-1}$ cannot be literally true. One major issue is that $\mathbf{L}_k^{-1}$ is a linear operator. Taken literally, linearity of the right-hand side of Equation (5.18) would require that every bit of information about the change in $f(t)$ is reflected, at least a little bit, in $\tilde{f}(\overset{*}{\tau})$, which seems unreasonable. Another

serious problem is that empirical values of k estimated from neural data can be quite high (Cao *et al.*, 2021). This is a computational problem in that computing the kth derivative becomes more and more sensitive to noise as k increases (Shankar & Howard, 2012). In real cortical circuits, recurrent feedback involving networks of inhibitory interneurons works to dampen noise (Ferster & Miller, 2000). Nonetheless, $\mathbf{L}_k^{-1}$ captures some important phenomena of neural firing that should be taken seriously. First, the weights of $\mathbf{L}_k^{-1}$ do not reflect any type of learning or experience with the stimuli. They only extract information embedded in a population with different decay rates. Second, the shape of the receptive fields $\mathbf{L}_k^{-1}$ predicts for $\tilde{f}$ seems to agree reasonably well with experiment (Howard *et al.*, 2014), at least in cases with a few discrete stimuli presented widely separated in time. Third, the idea of using derivatives with respect to s as a signal to infer the time of a stimulus presentation is a sound idea, even if the brain doesn't literally use the Post approximation with $k = 38$ (or some other very large value of k) to extract this information.

A Logarithmic Scale for Past Time

Note that although Equation (5.23) is written as an integral transform of $f_t(\tau)$, it is not necessary to retain a detailed memory of $f_t(\tau)$. Updating Equation (5.16) requires only the preceding value $F_{t-dt}(s)$ and the momentary value $f(t)$; there is no need to retain prior values of f above and beyond the information present in $F_t(s)$. Moreover, $\tilde{f}_t(\overset{*}{\tau})$ can be computed from $F_t(s)$. We thus have a choice to make about how much information to retain in $F_t(s)$. That is, the brain can't actually have an infinite number of values of s. And there is no reason *a priori* to assume that the s values that are sampled should be evenly spaced. Because $\overset{*}{\tau} \equiv k/s$, choosing how to distribute the s also specifies how to distribute the $\overset{*}{\tau}$. Equations (5.23) and (5.24) suggest a specific choice for sampling $\overset{*}{\tau}$.

Consider $\tilde{f}$ at two nearby values of $\overset{*}{\tau}$, which we'll refer to as $\overset{*}{\tau}_o$ and $\overset{*}{\tau}_o + \epsilon$. If we observe $\tilde{f}_t(\overset{*}{\tau}_o)$ and find that it is at a high value, we know that $\tilde{f}_t(\overset{*}{\tau}_o + \epsilon)$ is also likely to be at a high value. Conversely, if we observe that $\tilde{f}_t(\overset{*}{\tau}_o)$ is close to zero, we know that $\tilde{f}_t(\overset{*}{\tau}_o + \epsilon)$ is also likely to be close to zero. Because they are affected by nearby points in time, these two values of $\tilde{f}$ are correlated with one another. Each value of $\overset{*}{\tau}$ we sample costs us something (e.g., metabolic energy for a brain, availability of RAM in a computer simulation, etc.). In the limit as $\epsilon \to 0$, there is no benefit to measuring $\tilde{f}$ at a second value. As ϵ increases from zero, the two values of $\tilde{f}$ provide different information about the past and there is some benefit to counteract the cost of sampling a second value of $\overset{*}{\tau}$. However, the benefit from a particular number ϵ depends on the choice of the first $\overset{*}{\tau}$. To get an intuition into why this is so, suppose that we start with a specific $\overset{*}{\tau}$ and specific ϵ, then we vary $\overset{*}{\tau}$ while keeping ϵ fixed. As we increase $\overset{*}{\tau}$, the impact of a fixed value of ϵ becomes less and less. This is true because Φ in Equation (5.23) depends only on the ratio

$\frac{\tau}{\overset{*}{\tau}}$ and the difference between $\frac{\tau}{\overset{*}{\tau}}$ and $\frac{\tau}{\overset{*}{\tau}+\epsilon}$ grows smaller as $\overset{*}{\tau}$ increases for all τ. If we adopt the strategy of choosing ϵ so that each additional value of $\overset{*}{\tau}$ provides the same benefit, we arrive at a sampling strategy where the difference between adjacent values of $\overset{*}{\tau}$ goes up linearly with the value $\overset{*}{\tau}$. One can formalize this further.[6]

Setting the spacing between adjacent samples of $\overset{*}{\tau}$ to be proportional to the starting value of $\overset{*}{\tau}$ leads immediately to several properties. First, the ratio between adjacent values must be a constant:

$$\overset{*}{\tau}_{n+1} - \overset{*}{\tau}_n = c\overset{*}{\tau}_n \implies \frac{\overset{*}{\tau}_{n+1}}{\overset{*}{\tau}_n} = 1 + c. \tag{5.25}$$

Second, the number of units one observes with a particular value of $\overset{*}{\tau}$ should go down with that value of $\overset{*}{\tau}$:

$$\frac{dn}{d\overset{*}{\tau}} = \frac{1}{\overset{*}{\tau}}. \tag{5.26}$$

This expression diverges at zero, which is obviously not physical. One solution is to fix some minimum value of $\overset{*}{\tau}$ that can be sampled, $\overset{*}{\tau}_{\text{min}}$.[7] Third, the samples of $\overset{*}{\tau}$ are evenly spaced as a function of the logarithm of $\overset{*}{\tau}$:

$$\overset{*}{\tau}_n = (1 + c)^n \overset{*}{\tau}_{\text{min}}, \tag{5.27}$$

$$n = \log_{1+c} \overset{*}{\tau}_n - \log_{1+c} \overset{*}{\tau}_{\text{min}}. \tag{5.28}$$

This cluster of properties is quite theoretically satisfying. Many sensory receptors in the mammalian brain sample continuous dimensions at logarithmically spaced intervals. For instance, the density of receptors on the retina has long been known to decrease linearly with distance from the center of the retina [as in Equation (5.26)], a property that appears to be respected throughout early stages of the visual system in the brain. Psychologically, the logarithmic sampling of time [Equation (5.28)] provides a close correspondence with the Weber–Fechner law from psychophysics, which states that the magnitude of a perceptual variable goes up linearly with the logarithm of the physical stimulus that causes it. The Weber–Fechner law holds (at least approximately over some range) for a number of simple stimulus dimensions (e.g., loudness of a tone, pitch of a tone, length of lines, etc.) and has been argued to hold for perception of temporal intervals as well. It would be quite elegant if the brain distributes receptors along a time axis using the same mathematical expression as receptors along the retina, resulting in similar perceptual properties. It is especially satisfying that the arguments leading to logarithmic distribution of "time receptors" made no reference to

6 For instance, it can be shown that if $\tilde{f}$ is driven by white noise, the mutual information between two values of $\tilde{f}$ sampled over time depends on the ratio of their $\overset{*}{\tau}$s (see Appendix A.1 of Shankar & Howard, 2013).

7 If it is important to sample zero, one could use some other sampling scheme for values below some threshold in order to arrive at zero (Howard & Shankar, 2018).

these data. Rather, Equations (5.25)–(5.28) were derived from a property of the Post approximation coupled with the argument that the brain ought to equalize redundancy among the receptors.

5.3.2 Behavioral Models Using Scale-Invariant Temporal History

The scale-invariant temporal history described in Section 5.3.1 can be used to construct a wide variety of behavioral models of memory. It is straightforward to extend temporal context models by using $\tilde{f}_t(\overset{*}{\tau})$ in place of $\mathbf{c}_t$. The primary result is that one obtains scale-invariant recency and contiguity effects (Figure 5.4). However, the temporal history $\tilde{f}_t(\overset{*}{\tau})$ can also be used to construct computational models of very different tasks that cannot be readily modeled using temporal context models. Some of these tasks are believed to rely on different "kinds of memory" than free recall.

Scale-Invariant Temporal Context Models

TCMs rely on the temporal autocorrelation of the temporal context vector in order to generate recency and contiguity effects – that is, even in a list of random words, the expectation of $\mathbf{c}_t^{\mathrm{T}}\mathbf{c}_{t+\mathrm{lag}}$ falls off gradually like ρ^{lag}. However, exponential functions set a strong scale. One can readily build a temporal context model using $\tilde{f}_t(\overset{*}{\tau})$ in place of $\mathbf{c}_t$. Rather than $\mathbf{M}^{CF}$ associating context vectors to items, one constructs an associative matrix for each $\overset{*}{\tau}$:

$$\frac{d\mathbf{M}(\overset{*}{\tau})}{dt} = f(t)\tilde{f}_t{}^{\mathrm{T}}(\overset{*}{\tau}). \tag{5.29}$$

Recall that $F(s)$ at a particular s is essentially a temporal context vector with $\rho = e^{-s}$. If one imagines $\mathbf{M}^{CF}(s)$ as the $\mathbf{M}^{CF}$ matrix one would get for each value of s as a function of s, then $\mathbf{M}(\overset{*}{\tau})$ is just that matrix-valued function of s, but with the inverse transform applied.[8] One may visualize $\mathbf{M}(\overset{*}{\tau})$ for a particular $\overset{*}{\tau}$ as a set of connections between a particular row in Figure 5.5b and the vector f. One obtains a probe as $\mathbf{f}^{\mathrm{IN}} \equiv \sum_n \mathbf{M}(\overset{*}{\tau}_n)\tilde{f}_p(\overset{*}{\tau}_n)$. Each list item is activated to the extent that the units in the temporal history when it was presented are also active in the probe. One may visualize this operation with respect to Figure 5.5b as follows. When a particular item is activated in $f(t)$, there is a particular pattern $\tilde{f}_t(\overset{*}{\tau})$. That item is activated according to the match between $\tilde{f}_t(\overset{*}{\tau})$ and the probe $\tilde{f}_p(\overset{*}{\tau})$, summing over rows (corresponding to the inner product) and columns (corresponding to the sum over $\overset{*}{\tau}_n$). In the case of a long list of nonrepeating words, it can be shown that this association falls off like a power-law function (Howard *et al.*, 2015). This property makes TCMs built in this way

8 The transform here would be applied from the right: $\mathbf{M}(\overset{*}{\tau}) = \mathbf{M}^{CF}(s)\left[\mathbf{L}_k^{-1}\right]^{\mathrm{T}}$.

scale-invariant. It is thus straightforward to build genuinely scale-invariant recency and contiguity effects.

TCMs built from a scale-invariant temporal history also have qualitatively different properties than TCMs that use only a single-scale temporal context vector. Consider a situation in which two items, A and B, are presented at a temporal separation of τ seconds. The temporal context for B has A presented τ seconds in the past. Let us repeat A and observe the prediction for B as A recedes into the past. First, in the case of a single temporal context vector, the temporal context for B is just $\rho^{\tau}\mathbf{c}_{\mathrm{A}}^{\mathrm{IN}}$. When A is repeated (neglecting retrieval of temporal context) it again contributes a $\mathbf{c}_{\mathrm{A}}^{\mathrm{IN}}$ term to the temporal context vector and B is cued by an amount proportional to ρ^{τ}. But now consider what happens in the time after A was repeated. In the time following repetition of A, the magnitude of the $\mathbf{c}_{\mathrm{A}}^{\mathrm{IN}}$ component of the temporal context vector decreases exponentially. As a consequence, B is cued less and less as A recedes into the past after its repetition. The behavior is very different if temporal context is constructed from $\tilde{f}(\overset{*}{\tau})$. As before, the temporal context that cues B is the representation of A presented τ seconds in the past. However, this corresponds to an $\tilde{f}$ in which units triggered by A with $\overset{*}{\tau}$ near τ are active. When A is repeated (again neglecting recovery of temporal context), it again triggers a sequence of cells. A time t after repeating A, the units with $\overset{*}{\tau}$ near t are active. But if $t \ll \tau$, these are different units than the ones that cue B. As the repetition of A recedes into the past, B is cued more as t approaches τ and then less as the sequence passes through the units that form the temporal context for B. Although the consequences of this property on models of free recall would be expected to be relatively subtle (there are many items composing the temporal context and retrieval of temporal context), this property could be extremely useful in other behavioral applications (e.g., serial recall).

Probing a Representation of What Happened When

The simple Hebbian model described in Section 5.1 is a special case of a class of distributed memory models called global match models. The name "global match" refers to the property that the probe is compared to one composite memory $\mathbf{M}$ that contains a mixture of information from all of the items in memory. Other distributed memory models made different assumptions. For instance, multitrace models (e.g., Hintzman, 1984; Shiffrin & Steyvers, 1997) assumed that memory is composed of a list of traces which can be selectively accessed based on the probes one provides as part of a query of memory. Each trace is a set of features stored at a particular time, closely analogous to $\mathbf{f}_t$ in the simple Hebbian model and TCMs.

The temporal context model sketched above using $\tilde{f}_p(\overset{*}{\tau})$ as a probe has the spirit of a global match model. One builds an associative $\mathbf{M}(\overset{*}{\tau}_n)$ and then takes a sum over both what and when information in constructing the output of memory, $\mathbf{f}^{\mathrm{IN}} = \sum_n \mathbf{M}(\overset{*}{\tau}_n)\tilde{f}_p(\overset{*}{\tau}_n)$. However, there are other ways one might query $\tilde{f}(\overset{*}{\tau})$ to construct behavioral models of different memory tasks. Multitrace models keep different elements of memory separate in a list. Because it maintains separable information

about what happened when, one can understand $\tilde{f}_t(\overset{*}{\tau})$ as a multitrace model, albeit one where the traces become more blurred together as time recedes into the past (Figure 5.5b). Behavioral modeling work has shown that by querying this representation in different ways, it's possible to construct quantitative behavioral models of different working memory tasks.

It is well established that people and animals can direct attention to a restricted region of visual space. Suppose that a participant maintains fixation at a particular spot in a visual display for a few seconds (in experiments a small spot is usually provided). Now suppose that the participant learns that something important will be presented in a particular region above and to the left of the location that is being fixated. It can be shown that the ability to perceive visual information is greater if a stimulus is presented in that region relative to a region where nothing in particular is expected. This increased perceptual and neural gain is referred to as "attention."

One can model attention, directed to particular regions of past time; this capability is important in constructing behavioral models of working memory tasks. Let us suppose that one can direct attention to particular regions of the timeline and then compute a vector-valued output like so:

$$\mathbf{f}^{\mathrm{o}} = \sum_n \tilde{f}(\overset{*}{\tau}_n) G(\overset{*}{\tau}_n). \tag{5.30}$$

Here, $G(\overset{*}{\tau})$ is an attentional weight that can highlight the contributions of items at different points in the past. It is not reasonable to suppose that attention can take the form of any arbitrary function over $\overset{*}{\tau}$. Let us suppose three constraints on the form of attention. First, attention can point at only one circumscribed region at a time. The function for attention should have one peak at a particular index n. Second, attention can be deployed over a wide region or a more narrow region depending on the task demands. To be concrete, given that attention is directed to a particular index n, one may imagine that the participant can control whether attention extends to many nearby indices, falling off gradually, or only extends to a few nearby indices, falling off more sharply. Notice that because of the spread in Φ over $\overset{*}{\tau}$ (e.g., see Figure 5.5b), even if attention was nonzero for exactly one index $\overset{*}{\tau}_n$, this would still allow information from nearby time points to contribute to $\mathbf{f}^{\mathrm{o}}$. These simple assumptions allow us to construct very different behavioral models from the same memory representation.

This flexibility is useful in modeling working memory tasks. Working memory is a term used to describe a form of memory that stores information with high precision for a short time. Working memory is an intellectual descendent of computational models based on STS and is believed to rely on brain regions distinct from the regions responsible for episodic memory tasks like delayed and continuous distractor free recall. The first of these working memory tasks is referred to as probe recognition; the second is judgment of recency (JOR). In both tasks, the participant is presented with a short list of highly memorable stimuli – to be concrete let's assume that the stimuli are letters of the alphabet presented visually on a computer screen. In both tasks, the lists are relatively short (say

10 items) and the memory test is given immediately. In both tasks, the stimuli are repeated many times over an experimental session lasting tens of minutes. In both tasks, the participant is given a probe consisting of letters for the memory test. The only (important) way the tasks differ is in the judgment the participant must make in response to the memory probe. In probe recognition, the participants' job is to press a button to indicate whether a probe stimulus was in the most recent list or not. Because the stimuli are repeated across many lists, the task is really to judge whether the probe was presented in a relatively broad region of time. In the short-term JOR task, participants are given a pair of probe stimuli and asked to select the probe stimulus that was presented more recently. Because both of the probe items came from the most recent list, short-term JOR requires more fine-grained judgments of the temporal record of the probe stimuli.

Although the details are beyond the scope of this chapter (Tiganj, Cruzado, & Howard, 2019), a careful study of accuracy and the amount of time it takes participants to respond shows that although both tasks show a robust recency effect, the manner in which memory is accessed is quite different. The findings from both experiments can be accommodated by models in which one makes a decision based on how well a probe overlaps with $\mathbf{f}^{\mathrm{o}}$, $\mathbf{f}_p^{\mathrm{T}}\mathbf{f}^{\mathrm{o}}$. The important difference between the model for probe recognition and JOR is how attention is deployed. In the model of probe recognition, attention is deployed broadly such that it's constant over the list. The overlap with the probe is thus stronger for more recent items and this strength falls off like a power law [Equation (5.23)]. This provides a respectable model of probe recognition (see especially Donkin & Nosofsky, 2012). In short-term JOR the pattern of results has long suggested that participants use what's called a self-terminating serial scanning model. We can build a serial scanning model over the scale-invariant temporal memory by supposing that the participant first sets attention to the recent present, such that only $G(\overset{*}{\tau}_1)$ is one. The participant then compares this output to the memory probes. After some very brief time, attention is shifted to a slightly less recent time point, for instance only $G(\overset{*}{\tau}_2)$ is nonzero. The decision terminates when a match is found. One can visualize this process with the help of Figure 5.5b. After studying the list G K L N T X H, suppose the correct answer is X. The participant will not find a match for X looking at the first several rows. The amount of time it takes to find a match and initiate a decision depends on how far in the past X was presented. If instead the correct answer was T, one would have to scan over a longer distance to find information about that probe, predicting a correspondingly longer response time. There are many more detailed quantitative predictions that follow from these models that can be worked out.

The important point here is that it is only possible to construct such distinct behavioral models because $\tilde{f}_t(\overset{*}{\tau})$ has separable information about what happened when. If the information about the time of past events was stored as a single number, as in the temporal context vector, it is much more difficult to imagine an attentional model, and certainly not one that aligns as well to our current understanding of visual attention.

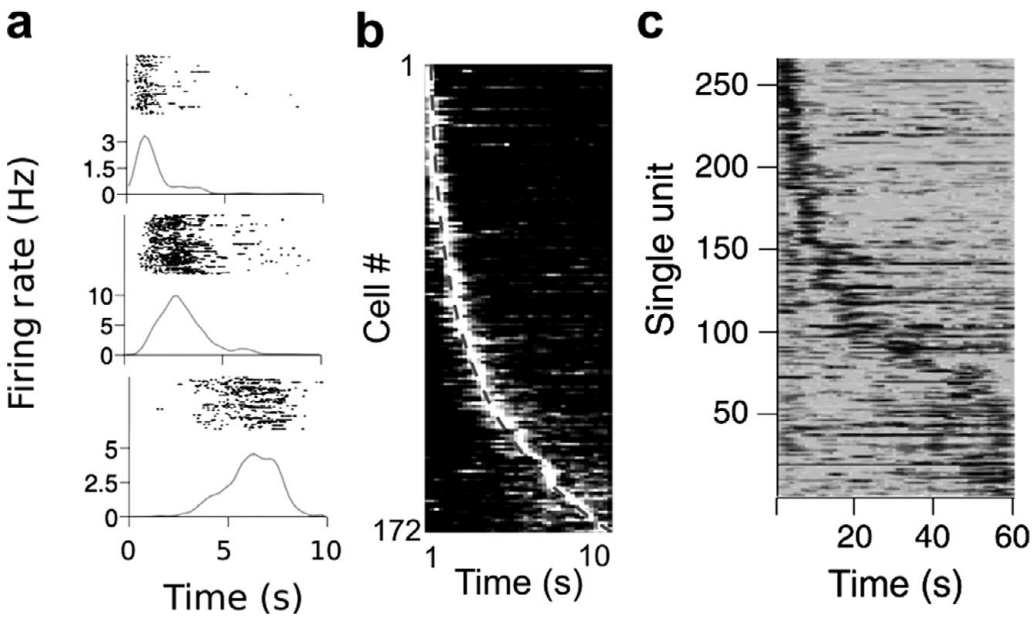


Figure 5.6 *So-called "time cells" are neurons that fire in sequence following a triggering stimulus. (a) Three time cells recorded from the hippocampus following the beginning of the delay period in a memory experiment. The top cell fires consistently over trials early in the delay. The middle and bottom cells also fire consistently, but at progressively later delays. After MacDonald* et al. *(2011). (b) A set of time cells in the hippocampus recorded during the delay interval sorted on their time of peak firing. Note that the population tiles the delay. This set of time cells could be used to determine the time within the delay. Note further that more cells fire earlier in the delay than later. This implies that there is greater resolution to the representation of time within the delay early in the delay period rather than later in the delay period. After Mau* et al. *(2018). (c) Time cells from the medial prefrontal cortex (mPFC). Note the scale of the x-axis extends out 60 s. After Bolkan* et al. *(2017).*

5.3.3 Evidence for Scale-Invariant Temporal History in the Brain

Taken literally, $\tilde{f}_t(\overset{*}{\tau})$ specifies the properties of a population of neurons. There is now extensive evidence for these predictions; populations of neurons referred to as "time cells" behave much as one might expect if they were implementing $\tilde{f}_t(\overset{*}{\tau})$ (see Figure 5.6). Let us take $\tilde{f}_t(\overset{*}{\tau})$ literally – as a description of the firing rate of a population of neurons, each indexed by a particular value of $\overset{*}{\tau}$. Time cells have now been observed in rodents (MacDonald *et al.*, 2011; Mello, Soares, & Paton, 2015; Pastalkova *et al.*, 2008; Tiganj *et al.*, 2017) and nonhuman primates (Cruzado *et al.*, 2020; Jin, Fujii, & Graybiel, 2009; Tiganj *et al.*, 2018) and been observed in studies in humans (Schonhaut *et al.*, 2022). Although the label "time cells" is most frequently applied to neurons in

the hippocampus, populations with similar properties have been observed in a variety of prefrontal regions as well as the striatum. These regions are believed to support different forms of memory. For instance, the hippocampus is believed to support episodic memory, prefrontal regions are believed to support working memory, and the striatum is believed to support implicit memory. If indeed different regions supporting different kinds of memory show firing consistent with properties of $\tilde{f}_t(\overset{*}{\tau})$, then this supports the hypothesis that behavioral models for different kinds of memory rely on the same form of representation.

Consider how cells representing $\tilde{f}_t(\overset{*}{\tau})$ would change their firing as a function of time following a delta function input at $t = 0$. Each cell would start with a firing rate near zero. As t approaches each cell's value of $\overset{*}{\tau}$, the firing rate of that cell would begin to increase, and then decrease again as t becomes much larger than that cell's $\overset{*}{\tau}$. Different cells have different values of $\overset{*}{\tau}$, so cells in the population would fire in sequence. The duration each cell spends firing depends linearly on its value of $\overset{*}{\tau}$; cells that fire later in the sequence should also fire for a longer time. Moreover, $\overset{*}{\tau}$s are sampled evenly over log rather than linear time, resulting in a decreasing number of cells that peak later in the sequence. Moreover, if the population carries information about what happened when, different stimuli should trigger distinguishable sequences. All of these properties have been quantitatively demonstrated in multiple brain regions, including the hippocampus and prefrontal regions in monkey and rodent. Moreover, time cells are observed in a wide variety of behavioral tasks (Cruzado *et al.*, 2020; Jin, Fujii, & Graybiel, 2009; MacDonald *et al.*, 2011; Mello, Soares, & Paton, 2015; Tiganj *et al.*, 2017, 2018), including in cases where the animal is given no task at all, but simply passively observes stimuli (Goh, 2021).

More recently, populations of neurons with properties like those predicted for $F_t(s)$ have been observed in a brain region called the entorhinal cortex (Bright *et al.*, 2020; Tsao *et al.*, 2018). Because they so closely resemble components of the temporal context vector [Equation (5.7)], these kinds of cells have been dubbed temporal context cells (see Figure 5.7). The entorhinal cortex provides the major projection to the hippocampus, where time cells were initially characterized. Decades of neurophysiology, neuropsychology, and cognitive neuroscience have implicated the entorhinal cortex and hippocampus in human episodic memory. For instance, the famous amnesia patient Henry Molaison (known prior to his death as H.M.) had bilateral damage to both the hippocampus and entorhinal cortex. Thus, a population of temporal context cells, which resemble $F_t(s)$, project to a population of time cells, which resemble $\tilde{f}_t(\overset{*}{\tau})$ in regions essential to human episodic memory.

5.3.4 Going Forward

The convergence between theoretical considerations (Section 5.3.1), behavioral models of memory (Section 5.3.2), and neurophysiological findings (Section 5.3.3)

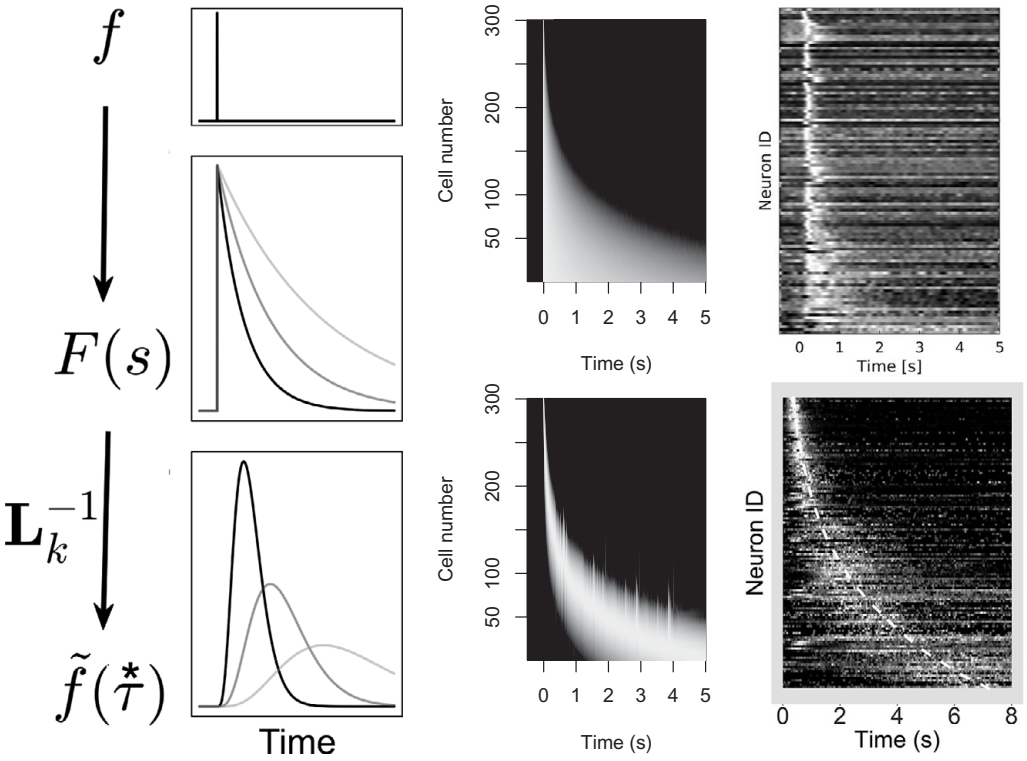


Figure 5.7 *Laplace transform of the past captures properties of temporal context cells and time cells. Left: Given a signal $f(t)$ as input, one can encode the real Laplace transform of the function leading up to the present using a bank of leaky integrators with rate constants s. Given a delta function input at time zero, each integrator in $F(s)$ rises to one and then decays exponentially. Each unit decays at a slightly different rate depending on that unit's value of s. The leaky integrators provide input to another population $\tilde{f}$ constructed by approximating the inverse Laplace transform via an operator $\mathbf{L}_k^{-1}$. Units in $\tilde{f}$ fire sequentially, with each cell peaking at a time controlled by the value of s that provides input to it. Middle: The two populations $F(s)$ (top) and $\tilde{f}$ (bottom) shown as heatmaps as a function of time to facilitate comparison with neurophysiological data. Right: These representations resemble so-called "temporal context cells" in the entorhinal cortex (top) and time cells in the hippocampus (bottom). Top after Bright* et al. *(2020). Bottom after Cao* et al. *(2021). Ian Bright and Rui Cao helped with this figure.*

seems very unlikely to happen by chance. This formalism could provide a foundation on which to build models of behavior and cognition that are more or less literal descriptions of the computations taking place in the brain. Although a foundation may exist, the work of constructing a complete theory of memory in the brain has barely begun. Thus far, the behavioral models that have been developed are sketches of important effects. A complete theory would require that these models be fleshed out to provide a detailed description of behavior (like the

models in Section 5.2). Development of such a theory would also require careful neuroscientific studies across species and tasks informed by these quantitative models of behavior. Theoretically, the formalism for encoding and inverting the Laplace transform of functions of time can be extended to representing functions over other variables. In this way it may prove possible to connect computational models of memory to well-developed computational models for spatial navigation, perception and simple decision-making informed by neurobiological data.

5.4 Related Literature

This chapter necessarily touched on only a tiny fraction of the data and computational models that have been used to understand human memory over the years. Kahana (2012) provides a thorough introduction to behavioral models of memory and important quantitative data from all the major human memory paradigms.

Stimulus-sampling theory is much more rich than decribed in this chapter. It was rigorously developed by many researchers, with Stanford University providing a focal point in the 1960s. Students interested in stimulus-sampling theory should consider the following papers: Atkinson and Estes (1962), Bower (1967).

Atkinson and Shiffrin (1968) is a modeling *tour de force* applying STS-based behavioral models to many variants of cued and free recall. It should be considered required reading for mathematical psychologists interested in modeling behavioral memory data. Raaijmakers and Shiffrin (1980) is a remarkably detailed desription of serial position effects in free recall that relies heavily on "fixed list context," an important concept in models of this era that is not discussed here (see also Criss & Shiffrin, 2005).

Howard (2018) provides a high-level review of cognitive and neural data related to the scale-invariant temporal history discussed in Section 5.3 (see also Howard & Hasselmo, 2020). Howard *et al.* (2015) built a number of simple cognitive models of behavioral tasks corresponding to different "kinds of memory" and note how this representation relates to distributed memory models. Lashley (1951) provides an eloquent critique of the limitations of simple associations in describing memory that seems to anticipate many of the properties of $\tilde{f}(\overset{*}{\tau})$ (see also James, 1890). There are also interesting connections between the logarithmic temporal scale derived for time here and measurement theory in mathematical psychology (for an overview, see Luce & Suppes, 2002) and exponential generalization (Shepard, 1987).

References

Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences*, *14*, 197–220.

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review*, *80*, 417–438.

Atkinson, R. C., & Estes, W. K. (1962). *Stimulus sampling theory* (No. 48). Citeseer.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–105). New York: Academic Press.

Baddeley, A. D., & Hitch, G. J. (1977). Recency reexamined. In S. Dornic (Ed.), *Attention and performance VI* (pp. 647–667). Hillsdale, NJ: Erlbaum.

Balsam, P. D., & Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. *Trends in Neuroscience*, *32*(2), 73–78.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6*, 173–189.

Bolkan, S. S., Stujenske, J. M., Parnaudeau, S., Spellman, T. J., Rauffenbart, C., Abbas, A. I., … Kellendonk, C. (2017). Thalamic projections sustain prefrontal activity during working memory maintenance. *Nature Neuroscience*, *20*(7), 987–996.

Bower, G. H. (1967). A multicomponent theory of the memory trace. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 1, pp. 229–325). New York: Academic Press.

Bright, I. M., Meister, M. L. R., Cruzado, N. A., Tiganj, Z., Buffalo, E. A., & Howard, M. W. (2020). A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proceedings of the National Academy of Sciences*, *117*, 20274–20283.

Bunsey, M., & Eichenbaum, H. B. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, *379*(6562), 255–257.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*, 313–323.

Cai, D. J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Song, W., … Silva, A. (2016). A shared neural ensemble links distinct contextual memories encoded close in time. *Nature*, *534*(7605), 115–118.

Cao, R., Bladon, J. H., Charczynski, S. J., Hasselmo, M., & Howard, M. (2021). Internally generated time in the rodent hippocampus is logarithmically compressed. *bioRxiv*, *2021.10.25.465750*.

Chan, S. C., Applegate, M. C., Morton, N. W., Polyn, S. M., & Norman, K. A. (2017). Lingering representations of stimuli influence recall organization. *Neuropsychologia*, *97*, 72–82.

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, *32*(1), 36–67. doi: 10.1080/03640210701801941

Criss, A. H., & Shiffrin, R. M. (2005). List discrimination in associative recognition and implications for representation. *Journal Experimental Psychology: Learning, Memory and Cogntion*, *31*(6), 1199–212. doi: 10.1037/0278-7393.31.6.1199

Cruzado, N. A., Tiganj, Z., Brincat, S. L., Miller, E. K., & Howard, M. W. (2020). Conjunctive representation of what and when in monkey hippocampus and lateral prefrontal cortex during an associative memory task. *Hippocampus*, *30*, 1332–1346.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*(1), 3–42.

Deitch, D., Rubin, A., & Ziv, Y. (2021). Representational drift in the mouse visual cortex. *Current Biology*, *31*(19), 4327–4339.

Donkin, C., & Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science*. doi: 10.1177/0956797611430961

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University.

Egorov, A. V., Hamam, B. N., Fransén, E., Hasselmo, M. E., & Alonso, A. A. (2002). Graded persistent activity in entorhinal cortex neurons. *Nature*, *420*(6912), 173–178.

Eichenbaum, H. (2017). On the integration of space, time, and memory. *Neuron*, *95*(5), 1007–1018. doi: 10.1016/j.neuron.2017.06.036

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94–107.

Estes, W. K. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*, 369–377.

Estes, W. K. (1955b). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*, 145–154.

Ferster, D., & Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience*, *23*(1), 441–471.

Folkerts, S., Rutishauser, U., & Howard, M. (2018). Human episodic memory retrieval is accompanied by a neural contiguity effect. *Journal of Neuroscience*, *38*, 4200–4211.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*(2), 289–344.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279–325.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67.

Glanzer, M. (1972). Storage mechanisms in recall. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (pp. 129–193). New York: Academic Press.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351–360.

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., & Gretz, A. L. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 355–369.

Goh, W. Z. (2021). *Remembering the past to predict the future: A scale-invariant timeline for memory and anticipation* (Unpublished doctoral dissertation). Boston University.

Goyal, A., Miller, J., Watrous, A. J., Lee, S. A., Coffey, T., Sperling, M. R., … Jacobs J. (2018). Electrical stimulation in hippocampus and entorhinal cortex impairs spatial and temporal memory. *Journal of Neuroscience*, *38*(19), 4471–4481.

Guo, C., Huson, V., Macosko, E. Z., & Regehr, W. G. (2021). Graded heterogeneity of metabotropic signaling underlies a continuum of cell-intrinsic temporal responses in unipolar brush cells. *Nature Communications*, *12*(1), 1–12.

Hasselmo, M. E., & McClelland, J. L. (1999). Neural models of memory. *Current Opinion in Neurobiology*, *9*, 184–188.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1–2), 1–34.

Healey, M. K., Long, N. M., & Kahana, M. J. (2018). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, 1–22.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers*, *16*(2), 96–101.

Hintzman, D. L. (1987). Recognition and recall in MINERVA 2: Analysis of the 'recognition-failure' paradigm. In P. Morris (Ed.), *Modelling cognition* (pp. 215–229). New York: Wiley.

Howard, M. W. (2018). Memory as perception of the past: Compressed time in mind and brain. *Trends in Cognitive Sciences*, *22*, 124–136.

Howard, M. W., & Hasselmo, M. E. (2020). Cognitive computation using neural representations of time and space in the laplace domain. *arXiv preprint arXiv:2003.11668*.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., & Eichenbaum, H. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *Journal of Neuroscience*, *34*(13), 4692–4707. doi: 10.1523/JNEUROSCI.5808-12.2014

Howard, M. W., & Shankar, K. H. (2018). Neural scaling laws for an uncertain world. *Psychologial Review*, *125*, 47–58. doi: 10.1037/rev0000081

Howard, M. W., Shankar, K. H., Aue, W., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*(1), 24–53.

Howard, M. W., Youker, T. E., & Venkatadass, V. (2008). The persistence of memory: Contiguity effects across several minutes. *Psychonomic Bulletin & Review*, *15*(PMC2493616), 58–63.

Hsieh, L.-T., Gruber, M. J., Jenkins, L. J., & Ranganath, C. (2014). Hippocampal activity patterns carry information about objects in temporal context. *Neuron*, *81*(5), 1165–1178.

Hull, C. L. (1939). The problem of stimulus equivalence in behavior theory. *Psychological Review*, *46*(1), 9.

Hull, C. L. (1947). The problem of primary stimulus generalization. *Psychological Review*, *54*, 120–134.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*, 208–233.

Hyman, J. M., Ma, L., Balaguer-Ballester, E., Durstewitz, D., & Seamans, J. K. (2012). Contextual encoding by ensembles of medial prefrontal cortex neurons. *Proceedings of the National Academy of Sciences USA*, *109*, 5086–5091. doi: 10.1073/pnas.1114415109

James, W. (1890). *The principles of psychology*. New York: Holt.

Jin, D. Z., Fujii, N., & Graybiel, A. M. (2009). Neural representation of time in cortico-basal ganglia circuits. *Proceedings of the National Academy of Sciences*, *106*(45), 19156–19161.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*, 103–109.

Kahana, M. J. (2012). *Foundations of human memory*. New York: Oxford University Press.

Killeen, P. R., & Fetterman, J. G. (1988). A behavioral theory of timing. *Psychological Review*, *95*(2), 274–295.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior; the Hixon Symposium* (pp. 112–146). Oxford: Wiley.

Lehman, M., & Malmberg, K. J. (2012). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*. doi: 10.1037/a0030851

Logan, G. D. (2021). Serial order in perception, memory, and action. *Psychological Review*, *128*(1), 1.

Luce, R. D., & Suppes, P. (2002). Representational measurement theory. In J. Wixted & H. Pashler (Eds.), *Stevens handbook of experimental psychology*, 3rd ed. (Vol. 4, pp. 1–41). Wiley Online Library.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737–749.

Mack, C. C., Cinel, C., Davies, N., Harding, M., & Ward, G. (2017). Serial position, output order, and list length effects for words presented on smartphones over very long intervals. *Journal of Memory and Language*, *97*, 61–80.

Mankin, E. A., Sparks, F. T., Slayyeh, B., Sutherland, R. J., Leutgeb, S., & Leutgeb, J. K. (2012). Neuronal code for extended time in the hippocampus. *Proceedings of the National Academy of Sciences*, *109*, 19462–19467. doi: 10.1073/pnas .1214107109

Manning, J. R., Polyn, S. M., Litt, B., Baltuch, G., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, *108*(31), 12893–12897.

Manns, J. R., Howard, M. W., & Eichenbaum, H. B. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, *56*(3), 530–540.

Mau, W., Hasselmo, M. E., & Cai, D. J. (2020). The brain in motion: How ensemble fluidity drives memory-updating and flexibility. *Elife*, *9*, e63550.

Mau, W., Sullivan, D. W., Kinsky, N. R., Hasselmo, M. E., Howard, M. W., & Eichenbaum, H. (2018). The same hippocampal CA1 population simultaneously codes temporal information over multiple timescales. *Current Biology*, *28*, 1499–1508.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

Mello, G. B., Soares, S., & Paton, J. J. (2015). A scalable population code for time in the striatum. *Current Biology*, *25*(9), 1113–1122.

Metcalfe, J. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review*, *92*, 1–38.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*(2), 839–862.

Nielson, D. M., Smith, T. A., Sreekumar, V., Dennis, S., & Sederberg, P. B. (2015). Human hippocampus represents space and time during retrieval of real-world memories. *Proceedings of the National Academy of Sciences*, *112*(35), 11078–11083.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

Palombo, D. J., Di Lascio, J. M., Howard, M. W., & Verfaellie, M. (2019). Medial temporal lobe amnesia is associated with a deficit in recovering temporal context. *Journal of Cognitive Neuroscience*, *31*(2), 236–248.

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, *321*(5894), 1322–1327.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*, 129–156.

Post, E. (1930). Generalized differentiation. *Transactions of the American Mathematical Society*, *32*, 723–781.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, *17*, 132–138.

Quenon, L., de Xivry, J.-J. O., Hanseeuw, B., & Ivanoiu, A. (2015). Investigating associative learning effects in patients with prodromal Alzheimer's disease using the temporal context model. *Journal of the International Neuropsychological Society*, *21*(09), 699–708.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207–262). New York: Academic Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectivenesss of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory.* New York: Appleton-Century-Crofts.

Rubin, A., Geva, N., Sheintuch, L., & Ziv, Y. (2015). Hippocampal ensemble dynamics timestamp events in long-term memory. *eLife*, *4*, e12247.

Rule, M. E., Loback, A. R., Raman, D. V., Driscoll, L. N., Harvey, C. D., & O'Leary, T. (2020). Stable task information from an unstable neural population. *Elife*, *9*, e51121.

Rule, M. E., O'Leary, T., & Harvey, C. D. (2019). Causes and consequences of representational drift. *Current Opinion in Neurobiology*, *58*, 141–147.

Schonhaut, D. R., Aghajan, Z. M., Kahana, M. J., & Fried, I. (2022). A neural code for spatiotemporal context. bioRxiv, https://doi.org/10.1101/2022.05.10.491339.

Schoonover, C. E., Ohashi, S. N., Axel, R., & Fink, A. J. (2021). Representational drift in primary olfactory cortex. *Nature*, 1–6.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*, 893–912.

Shankar, K. H., & Howard, M. W. (2010). Timing using temporal context. *Brain Research*, *1365*, 3–17.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, *24*(1), 134–193.

Shankar, K. H., & Howard, M. W. (2013). Optimally fuzzy temporal memory. *Journal of Machine Learning Research*, *14*, 3753–3780.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM — retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*, 145–166.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135–171.

Talamonti, D., Koscik, R., Johnson, S., & Bruno, D. (2021). Temporal contiguity and ageing: The role of memory organization in cognitive decline. *Journal of Neuropsychology*, *15*, 53–65.

Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., & Howard, M. W. (2018). Compressed timeline of recent experience in monkey lPFC. *Journal of Cognitive Neuroscience*, *30*, 935–950.

Tiganj, Z., Cruzado, N. A., & Howard, M. W. (2019). Towards a neural-level cognitive architecture: Modeling behavior in working memory tasks with neurons. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1118–1123). Montreal: Cognitive Science Society.

Tiganj, Z., Hasselmo, M. E., & Howard, M. W. (2015). A simple biophysically plausible model for long time constants in single neurons. *Hippocampus*, *25*(1), 27–37.

Tiganj, Z., Kim, J., Jung, M. W., & Howard, M. W. (2017). Sequential firing codes for time in rodent mPFC. *Cerebral Cortex*, *27*, 5663–5671.

Trutti, A. C., Verschooren, S., Forstmann, B. U., & Boag, R. J. (2021). Understanding subprocesses of working memory through the lens of model-based cognitive neuroscience. *Current Opinion in Behavioral Sciences*, *38*, 57–65.

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., & Moser, E. I. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature*, *561*, 57–62.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford.

Uitvlugt, M. G., & Healey, M. K. (2019). Temporal proximity links unrelated news events in memory. *Psychological Science*, *30*(1), 92–104.

Yaffe, R. B., Kerr, M. S. D., Damera, S., Sarma, S. V., Inati, S. K., & Zaghloul, K. A. (2014). Reinstatement of distributed cortical oscillations occurs with precise

spatiotemporal dynamics during successful memory retrieval. *Proceedings of the National Academy of Sciences*, *111*(52), 18727–18732. doi: 10.1073/pnas.1417017112

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179.

6 Statistical Decision Theory

F. Gregory Ashby and Michael J. Wenger

6.1 Introduction

In 2002, Estes referred to signal detection theory (SDT) as "the most towering achievement of basic psychological research in the last half century" (p. 15). SDT is, by far, the most dominant model in psychophysics, and its multidimensional generalization has become the default approach for defining and studying perceptual interactions. The name "signal detection theory" refers to applications of the theory to tasks in which only one stimulus dimension is relevant, and the most common version requires participants to detect a signal embedded in noise. Tasks that require attention to more than one stimulus dimension typically require a decision more complex than simple detection – for example, the participant may be required to identify the presented stimulus uniquely, or assign it to a predetermined category. In such cases, the same statistical model is more appropriately called general recognition theory (GRT). We refer to both approaches by the term *statistical decision theory*.

This chapter reviews statistical decision theory, beginning with its origins, laying out its foundations in one dimension and its extension to two or more dimensions. We describe applications of the theory to identification and classification tasks, to the perception of configurality and holism, to the modeling of response times (RTs), and finally we consider extensions to neuroscience. An overarching theme of this chapter is that statistical decision theory provides a consistently evolving, general and powerful approach to modeling decision processes involved in sensation, perception, and cognition.

6.2 Historical Precedents

Statistical decision theory emerged when two simple propositions were applied to a new experimental paradigm that eventually formed the foundation of psychophysics and much of experimental psychology. The first of these is the proposition that one can experience the qualia of a known stimulus (such as light) even in the absence of that stimulus. Perhaps the most famous example of this is the Helmholtz (1867) thought experiment on phosphenes: mechanical pressure on the eye causes the subjective experience of patterns of light even in a dark room.[1] Similarly, one can fail to experience the qualia of a known stimulus even when that stimulus is present (e.g., a light is on, but may be too dim to see). It appears that thinking about such possibilities was at the root of the classic two-alternative forced-choice design, and that thoughts about these possibilities are evident in work by both Fechner and Thurstone (Fechner, 1860; Link, 1994; Wixted, 2020).

The second of the two simple propositions is the idea that encoded psychological information may be a combination of a fixed value and random error. The formal notion of this possibility in human measurement can be traced at least to the work of Gauss (Dunnington, Gray, & Dohse, 2004), and the general notion of random variation in subjective human experience dates at least to the work of Cattell (Fullerton & Cattell, 1892) and Thurstone (Thurstone, 1927a, 1927b). However, the formal treatment of randomness in support of decision-making, as it has come to be expressed in statistical decision theory, emerged from the (at times contentious) debates that Fisher had with Neyman and Pearson (Fisher, 1955; Neyman & Pearson, 1933). In particular, Neyman and Pearson's distinction between Type I and Type II errors – corresponding to false alarms and misses, respectively – was offered as a refinement to Fisher's notion of a *p*-value, which itself had originally been proposed as an objective, though informal, index of the level of trust in a null hypothesis (Lenhard, 2006).

The initial linking of these two simple propositions occurred in early work on radar and sonar and other areas of electronics and electrical engineering. It appears that the basic vocabulary of SDT – hits, misses, false alarms, and correct rejections – emerged from the World War II need, for example, to determine whether to

1 Curiosity about phosphenes predates Helmholtz, as sketches of phosphenes can be found in Newton's notes (http://cudl.lib.cam.ac.uk/view/MS-ADD-0397).

drop a bomb or a depth charge on a possible enemy submarine (Marcum, 1947). Likewise, as noted by Wixted (2020), the idea that the probabilistic behavior of photographic film and television tubes might provide a model for the human visual system had already been considered in electrical engineering (Rose, 1942, 1948). The explicit merging of these ideas and their application to the analysis of both human behavior and the performance of engineered systems appears to have occurred at about the same time at MIT and the University of Michigan (Creelman, 2015; Peterson & Birdsall, 1953; Peterson, Birdsall, & Fox, 1954; Van Meter & Middleton, 1954).

In each of these contexts, the canonical experiment includes trials in which a stimulus or signal is or is not present and the observer or system is required to respond that the signal is present or absent. This task inspired the name "signal detection theory," and almost all modern applications of SDT are either to this task or to the logically equivalent two-stimulus identification task, which we consider in detail in the next section. In fact, Link (1994) rightly noted that the use of this canonical task goes back at least to Fechner's foundational work on psychophysics. As we will see, this simple experiment provides a powerful and general framework for understanding how signals are processed – either by biological or engineered systems.

To illustrate the power and generality of this accomplishment (and to reflect on Estes' evaluation), we obtained rough estimates of the number of publications that used SDT in audition and vision, in 5-year increments between 1955 and early 2020.[2] We contrasted these data with the number of PhDs awarded in experimental, cognitive, and human factors psychology, along with the number of PhDs awarded in electrical, electronics, and communications engineering for that same range of years.[3]

Figure 6.1a plots the cumulative number of publications in audition and vision that include SDT, along with the number of PhDs awarded in psychology and engineering. This presentation is somewhat misleading, so Figure 6.1b plots the same data in terms of relative cumulative number (i.e., dividing the value of each data series at time t by the value at the starting point, 1960). It becomes apparent that the increase in the use of SDT is not simply due to an increase in the number of scientists who could potentially use SDT. This powerfully underscores Estes' estimate of SDT as a towering achievement. With this historical context in mind, we now consider the details.

2 Searches were performed using Google Scholar. The search for publications in audition was performed using "auditory OR audition OR perception signal detection theory -vision -visual" and the search in vision was performed using "vision OR visual OR perception signal detection theory -auditory -audition."

3 National Science Foundation, National Center for Science and Engineering Statistics, Survey of Earned Doctorates.

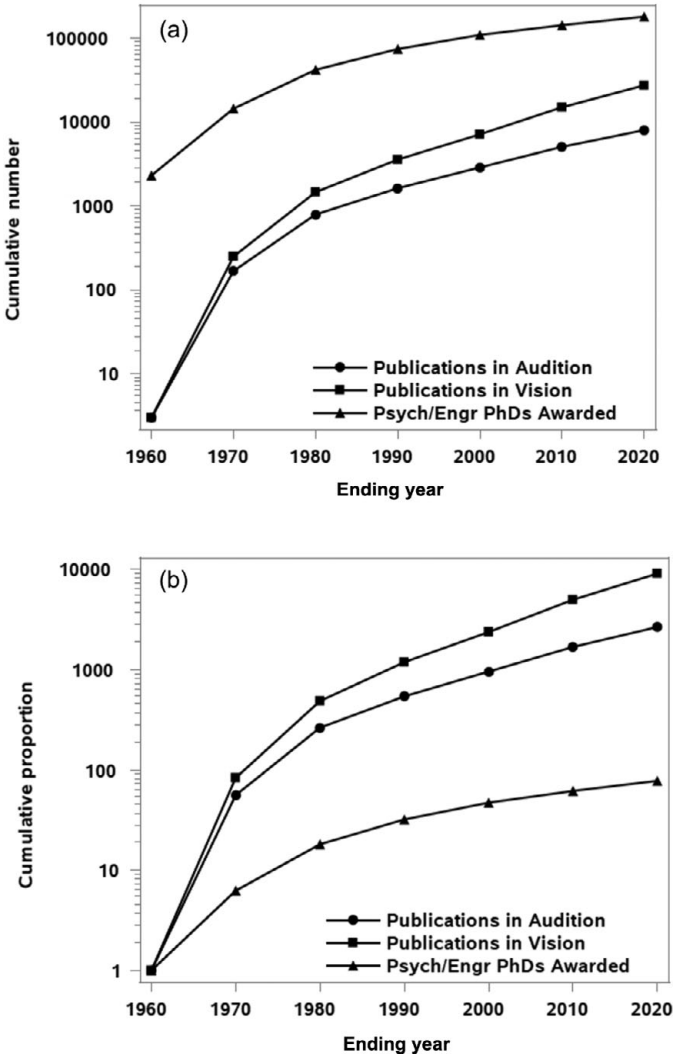


Figure 6.1 *(a) Cumulative publications citing signal detection theory in audition and vision, relative to cumulative PhDs awarded in sub-disciplines of psychology and engineering, 1960–2020. (b) Relative increase in publications citing signal detection theory in audition and vision, and relative increase in PhDs awarded in sub-disciplines of psychology and engineering, 1960–2020.*

6.3 One Dimension: Signal Detection Theory

The most common application of SDT is to a two-stimulus identification task – that is, a task with two stimuli and two uniquely identifying responses.[4]

4 See Macmillan and Creelman (2005) for an excellent comprehensive treatment of the practicalities of using signal detection theory.

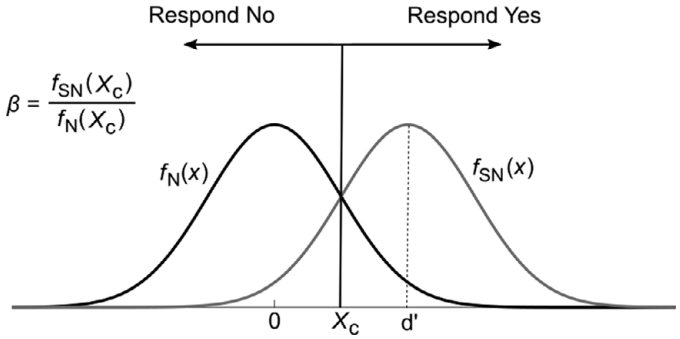


Figure 6.2 *The normal, equal-variance, SDT model.*

On each trial, the observer's task is to identify the single presented stimulus by emitting the appropriate response. In the original applications, the two stimuli were pure noise (N) and a signal of some type embedded in noise (SN). The observer's task was to indicate whether or not a signal was presented by responding YES or NO.

The standard SDT model for this YES–NO detection task is illustrated in Figure 6.2. The model assumes that performance in this task is based on a single sensory value, denoted by **X**. As described earlier, a fundamental assumption is that all sensations are inherently noisy, and thus **X** is a random variable. In the YES–NO detection task where the stimuli are N and SN, **X** represents sensory magnitude – for example, loudness with auditory stimuli, or brightness with visual stimuli. The probability density function (pdf) describing the distribution of sensory values on N trials is denoted by $f_N(x)$ and $f_{SN}(x)$ describes this distribution on SN trials. In Figure 6.2, both of these distributions are normal with the same variance. This normal, equal-variance model is the most commonly used model in signal detection analysis, but any distributions are possible.

Another fundamental assumption of SDT is that there is no fixed threshold on sensation that determines whether or not an observer will detect a signal. Instead, the observer is assumed to set a criterion value, denoted by X_C, and then use the following decision rule:

$$\text{Respond YES if } \mathbf{X} > X_C; \text{ otherwise, respond NO.} \tag{6.1}$$

Unlike the classical notion of a fixed threshold, the SDT criterion is under the observer's control. The observer is assumed to choose the value of X_C in a way that is typically assumed to depend on the costs of the two types of errors (i.e., misses and false alarms), the benefits of the two types of correct responses (i.e., hits and correct rejections), and on the N and SN base rates. Thus, in SDT, control of the criterion is relegated to decision processes, whereas the classical account assumed a fixed threshold for sensation that was a feature of sensory systems.

The response accuracy data are typically reported in a confusion matrix that includes a row for every stimulus and a column for each response. The entry in row i and column j is the number of stimulus i trials for which the observer responded j. When there are only two stimuli and two responses, then the confusion matrix is 2×2. The entries in row i add to the number of stimulus i trials in the experiment, and therefore do not depend on the data. As a result, each row includes only one degree of freedom (i.e., only one independent data value), so no information is lost if only one entry in each row is reported. The standard is to report the entries in the column associated with the YES response. These are used to estimate the probability of a false alarm (i.e., responding YES on N trials) and the probability of a hit (responding YES on SN trials). From Figure 6.2 it is easily seen that

$$P(\text{FA}) = 1 - F_{\text{N}}(X_{\text{C}}), \tag{6.2}$$

where $F_{\text{N}}(X_{\text{C}})$ is the cumulative distribution function of the N distribution, evaluated at X_{C}. Similarly:

$$P(\text{H}) = 1 - F_{\text{SN}}(X_{\text{C}}). \tag{6.3}$$

In any two-stimulus identification task, the data have two degrees of freedom [e.g., $P(\text{H})$ and $P(\text{FA})$]. The SDT model shown in Figure 6.2 has two free parameters – the location of the response criterion, denoted by X_{C}, and the distance between the means of the N and SN distributions in standard deviation units, denoted by d'. If the normal, equal-variance model is assumed, then X_{C} and d' can be estimated by inverting Equations (6.2) and (6.3). Specifically, X_{C} is estimated by inverting Equation (6.2) to produce

$$\hat{X}_{\text{C}} = \Phi^{-1}\left[1 - \hat{P}(\text{FA})\right], \tag{6.4}$$

where Φ^{-1} is the inverse-Z transformation [i.e., $\Phi^{-1}(p)$ is the Z-value that has area to the left equal to p] and $\hat{P}(\text{FA})$ is the observed proportion of false alarms. Note from Figure 6.2 that d' equals the standardized distance from the mean of the N distribution to X_{C} (i.e., X_{C}) plus the distance from X_{C} to the mean of the SN distribution. Therefore

$$\hat{d}' = \hat{X}_{\text{C}} - \Phi^{-1}\left[1 - \hat{P}(\text{H})\right]. \tag{6.5}$$

Note also that d' is the standardized distance between the means (i.e., the mean difference divided by the common standard deviation). As a result, the common variance is not identifiable, in the sense that any combination of mean differences and standard deviations that combine to produce the same d' will make identical predictions. As a result, we can set the common standard deviation to 1 without loss of generality.

The two degrees of freedom in the data can be used to estimate X_{C} and d', but then there are no data left to test the model's goodness-of-fit. Given that the model

can perfectly fit any observed values of $P(\mathrm{H})$ and $P(\mathrm{FA})$, an obvious question is why fit this model to two-stimulus identification data? The most common reason, which has been confirmed in thousands of applications, is that SDT is highly successful at separating perceptual and decisional effects. In particular, manipulations that should only affect sensory magnitude – such as increasing or decreasing signal intensity – mostly cause d' to change but not X_{C}, whereas manipulations that should only affect the observer's decision about how to act on their sensory experience – such as changing the costs and benefits associated with the various possible outcomes – mostly cause X_{C} to change but not d'. In contrast, any of these changes are likely to cause accuracy to change, so without SDT, it is generally impossible to know whether a change in accuracy is due to a change in perception or a change in decision strategy. SDT offers a highly effective method for solving this problem.

6.3.1 The Receiver Operating Characteristic

A standard way to summarize the results of a YES–NO detection experiment is via the receiver operating characteristic (ROC), which plots $P(\mathrm{H})$ (on the ordinate) against the probability of a false alarm $P(\mathrm{FA})$ (on the abscissa). The standard approach is to plot data from a variety of conditions that cause X_{C} to change, but not d'. Examples are shown in Figure 6.3. Because each point on any one curve is associated with a different value of X_{C} but the same value of d', these are iso-sensitivity contours. Technically, other kinds of curves could be plotted in the same space (e.g., iso-bias curves), but because iso-sensitivity contours are so common, this is almost always what is meant by an ROC curve. For any positive value of d', the iso-sensitivity curve must fall completely in the upper left half of the plot. The main diagonal, in which $P(\mathrm{H}) = P(\mathrm{FA})$ (denoted by the dotted line), corresponds to $d' = 0$. Any curve (or point) below this diagonal indicates a negative d'. Since pure guessing should produce $d' = 0$, a (significantly) negative d' should only occur because of participant deception or because the observer is using a highly suboptimal decision rule.

There are several popular experimental designs that are used to estimate iso-sensitivity curves. One approach is to include a variety of conditions in which the stimulus characteristics remain fixed, but different payoffs are used to encourage participants to change their criterion for responding YES. Another approach, which uses the same N and SN trials but is experimentally more efficient, is to ask observers to rate the intensity of the signal on each trial. Given an r-point rating scale, $r - 1$ points on an iso-sensitivity curve can be estimated by assuming that observers construct $r - 1$ criteria, denoted by $X_1, X_2, \ldots, X_{r-1}$, and respond with rating i if and only if $X_{i-1} < \mathbf{X} \leq X_i$, where $X_0 = -\infty$ and $X_r = \infty$. The ith point on the curve is then estimated via

$$\hat{P}(\mathrm{FA}_i) = \hat{P}(R > i|\mathrm{N}) \tag{6.6}$$

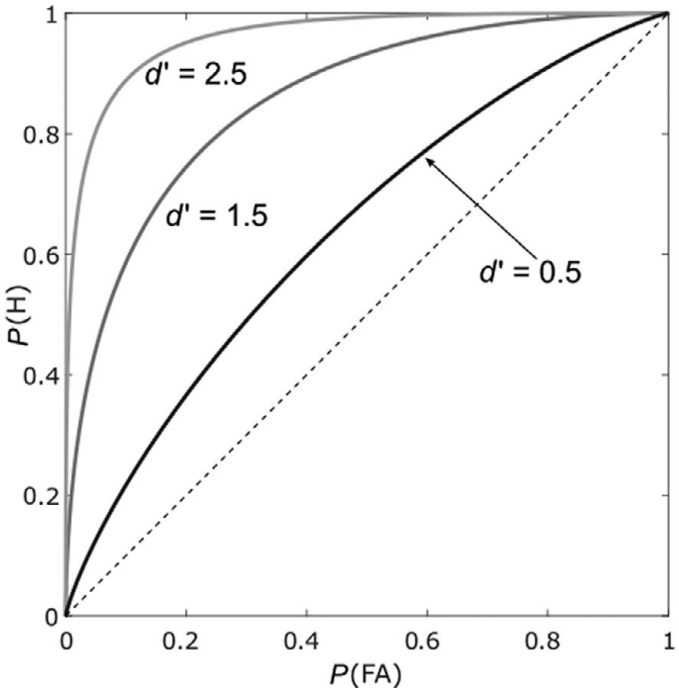


Figure 6.3 *An ROC showing iso-sensitivity contours for three different values of d'.*

and

$$\hat{P}(\mathrm{H}_i) = \hat{P}(R > i|\mathrm{SN}), \tag{6.7}$$

where R is the observer's rating.

The optimal decision strategy in any two-stimulus identification task depends on the likelihood ratio

$$L(x) = \frac{f_{\mathrm{SN}}(x)}{f_{\mathrm{N}}(x)}. \tag{6.8}$$

In particular, if the goal is to maximize the probability of a correct decision, then the optimal decision rule is to

$$\text{Respond YES if } L(\mathbf{X}) > \frac{P(\mathrm{N})}{P(\mathrm{SN})}\text{; otherwise, respond NO,} \tag{6.9}$$

where $P(\mathrm{N})$ and $P(\mathrm{SN})$ are the probabilities that N and SN, respectively, are presented on each trial (i.e., the stimulus base rates). Thus, if SN and N are equally likely, then the optimal strategy is to respond YES if the current sensory magnitude is more likely to be a sample from the SN distribution than from the N distribution. If the sample is more likely from the N distribution, then the NO response should be given. This is the scenario in Figure 6.2. If there are more N trials than SN trials,

then the Equation (6.9) decision rule indicates that stronger evidence is required before responding YES.

In some applications, the different types of errors may incur different penalties and the different types of correct decisions may bring different benefits. Let $V_{I,J}$ denote the value (either positive or negative) of responding J (e.g., YES or NO) on trials when stimulus I was presented (e.g., SN or N). Then the decision rule that maximizes value is (e.g., Green & Swets, 1966)

$$\text{Respond YES if } L(\mathbf{X}) > \frac{(V_{\text{N,NO}} + V_{\text{N,YES}})P(\text{N})}{(V_{\text{SN,YES}} + V_{\text{SN,NO}})P(\text{SN})}; \text{ otherwise, respond NO.} \tag{6.10}$$

Note that according to this rule, if the only change in the outcomes is to increase the reward for a correct rejection – that is to increase the (positive) value of $V_{\text{N,NO}}$ – then the observer should increase the criterion, since this will ensure more NO responses. In contrast, if the only change is to increase the penalty for a false alarm – that is to decrease the (negative) value of $V_{\text{N,YES}}$ – then the observer should decrease the criterion, since this will ensure fewer YES responses.

Because of the important role that the likelihood ratio plays in optimal responding, the Equation (6.1) decision rule is sometimes reformulated in terms of the likelihood ratio:

$$\text{Respond YES if } L(\mathbf{X}) > \beta; \text{ otherwise, respond NO.} \tag{6.11}$$

In this version of the theory, β can be interpreted as the value of the likelihood ratio at the criterion X_{C} – that is

$$\beta = L(X_{\text{C}}) = \frac{f_{\text{SN}}(X_{\text{C}})}{f_{\text{N}}(X_{\text{C}})}. \tag{6.12}$$

As with X_{C}, the criterion β is assumed to be under the observer's control. Setting $\beta = P(\text{N})/P(\text{SN})$ maximizes accuracy [i.e., see Equation (6.9)], but the observer is free to set β at some other value. For example, the optimal value of β must be learned, and during this learning process, suboptimal values of β are to be expected.

Note that the Equation (6.1) and Equation (6.11) decision rules are equivalent if the likelihood ratio increases monotonically with $\mathbf{X}$. The Equation (6.1) decision rule responds NO to any $\mathbf{X} < X_{\text{C}}$ and YES to any $\mathbf{X} > X_{\text{C}}$, but if the likelihood ratio increases monotonically with $\mathbf{X}$, then the likelihood ratio is less than β for any $\mathbf{X} < X_{\text{C}}$ and greater than β for any $\mathbf{X} > X_{\text{C}}$, so under these conditions, the two decision rules always give the same response. This raises the obvious question of how one could tell from empirical data whether the likelihood ratio of the SN and N sensory distributions is or is not monotonically increasing with sensory magnitude. The key to answering this question is provided by the following result.

Theorem 6.1. *For any differentiable ROC curve, the likelihood ratio*

$$L(x) = \frac{f_{SN}(x)}{f_N(x)} \tag{6.13}$$

is a monotonically increasing function of x (i.e., sensory magnitude) if and only if the ROC is concave down.

Proof. By definition, a differentiable function is concave down if and only if its slope is monotonically decreasing. The slope of the ROC curve is

$$\frac{dP(\mathrm{H})}{dP(\mathrm{FA})} = \frac{d[1 - F_{\mathrm{SN}}(x)]}{d[1 - F_{\mathrm{N}}(x)]} = \frac{f_{\mathrm{SN}}(x)}{f_{\mathrm{N}}(x)} = L(x). \tag{6.14}$$

Therefore, the slope of the ROC curve equals the likelihood ratio, which proves the theorem. □

If the likelihood ratio increases monotonically with sensory magnitude, then the more intense the sensation, the greater the confidence that a signal was presented (i.e., SN). This makes sense, so we would expect empirical ROCs to be concave down, and in fact, the evidence strongly supports this prediction (Green & Swets, 1966). In other words, the empirical evidence supports the assumption that the likelihood ratio of the SN and N sensory distributions increases monotonically with sensory magnitude. These data rule out many alternative models of the N and SN distributions in which the likelihood ratio is not monotonic. Perhaps the best-known model in this class is the normal, unequal-variance model, which is illustrated in Figure 6.4. The top panel shows an N distribution with small variance and two alternative SN distributions, both with larger variances. The bottom panel shows the ROC curves predicted by this model under the assumption that the observer uses the Equation (6.1) decision rule.

Figure 6.4 displays several features worth noting. First, the likelihood ratio is not monotonically increasing. Note that, as expected, the SN distribution has higher likelihood for large sensory magnitudes, but non-intuitively, it also has higher likelihood for small magnitudes (i.e., magnitudes below the mean of the N distribution). Therefore, as sensory magnitude increases, the likelihood ratio is initially large (i.e., greater than 1), is then small (less than 1), and finally becomes large again (greater than 1). Because of this non-monotonicity, the Equation (6.1) decision rule is not optimal. Instead, the optimal strategy [i.e., described by Equation (6.11)] is to respond YES to small and large sensory magnitudes [when $L(\mathbf{X}) > 1$] and NO only for magnitudes of intermediate value [when $L(\mathbf{X}) < 1$].

Second, note that the ROC curves shown in Figure 6.4B are not concave down. Instead, the upper right portion of both curves displays a pronounced violation of concavity. Furthermore, note that both ROCs dip below the main diagonal, which, as mentioned earlier, reflects suboptimal decision making. This is because the predicted ROC curves shown in Figure 6.4 were generated under the assumption that the observer is using the Equation (6.1) decision rule, which is highly suboptimal for small sensory magnitudes.

Third, neither ROC curve in Figure 6.4B is symmetric around the negative diagonal. In fact, many empirical ROCs, albeit concave down, are skewed in this same manner (Green & Swets, 1966), and this is the main reason that the normal,

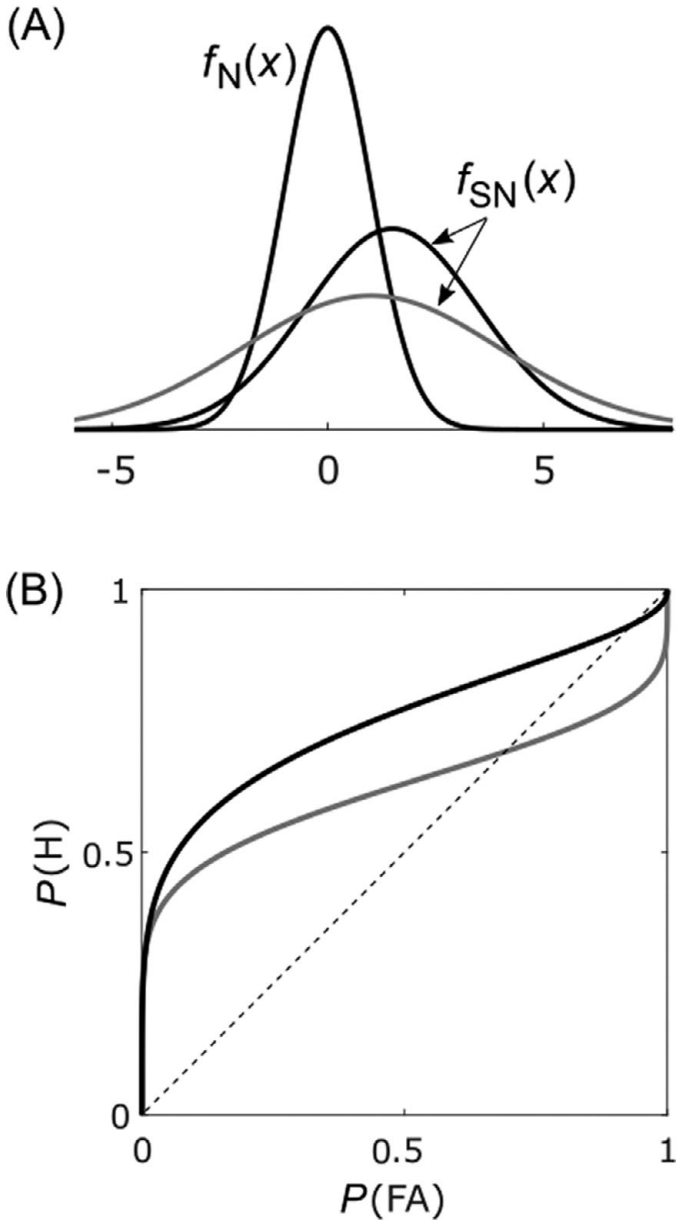


Figure 6.4 *(A) The normal, unequal-variance model of SDT. The N distribution is normal with mean 0 and variance 1. Two alternative SN distributions are shown. The pdf in black is normal with mean 1.5 and standard deviation 2, whereas the pdf in gray is normal with mean 1 and standard deviation 3. (B) The ROC showing the iso-sensitivity contours predicted by the two models shown in panel A. Both curves assume the N distribution is normal with mean 0 and variance 1. The black curve assumes the SN distribution has mean 1.5 and standard deviation 2, whereas the gray curve assumes the SN distribution has mean 1 and standard deviation 3.*

unequal-variance model is popular. In other words, this model accounts for the many reports that empirical ROC curves are skewed, but it is inconsistent with the ubiquitous finding that empirical ROCs are concave down.

Finally, note that the standard measure of sensitivity, namely d', is not defined in this model. Traditionally, d' is defined as the distance between the N and SN means divided by the common standard deviation. In the normal, unequal-variance model, however, there is no common standard deviation, so the traditional d' is undefined. This is also a common problem with multivariate extensions of SDT.

In summary, empirical ROC curves are concave down and are either approximately symmetric about the negative diagonal or skewed in the direction shown in Figure 6.4. The normal, equal-variance model accounts for symmetric ROCs that are concave down, but as it turns out, so do many other models. Killeen and Taylor (2004) describe the necessary conditions on the N and SN distributions for an SDT model to predict symmetric ROCs.[5] In addition, many SDT models account for skewed ROCs that are concave down. Included in this list, for example, are models in which the N and SN distributions are both exponential or Rayleigh distributions.

6.3.2 Application to Other Tasks

Although the original applications of SDT in psychology were to YES–NO detection tasks, the theory has also been applied to a variety of other tasks. First, applications to any two-stimulus identification task are identical except for relabeling of the stimuli and responses. For example, suppose the stimuli are "A" and "B" and their identifying responses are "a" and "b." If A and B are different stimuli then they must differ in some way. If they differ on some quantitative (i.e., prothetic) dimension, then associate the stimulus with the smaller value with N and its associated response with NO. If they differ on some qualitative (i.e., metathetic) dimension, then the association of A and B to N and SN is arbitrary. Either way, once the associations are complete, the SDT model is identical to the model for the YES–NO detection task.

In addition, SDT has been applied to a variety of different types of experiments that include multiple stimuli. The most widely used is probably the two-sample, two-alternative forced-choice task. On each trial, two stimuli are presented – one N and one SN (or one A and one B), and the observer's task is to identify which one is SN (or e.g., B). SDT assumes that exposure to the two stimuli produces two sensory magnitudes – one that is a random sample from the N distribution and one randomly sampled from the SN distribution – and that the observer identifies the larger of these as SN. A well-known result, described in the following proposition, is that the probability correct in this task equals the area under the ROC that results from the YES–NO detection task (Green, 1964; Green & Swets, 1966).

5 Specifically, the ROC is symmetric if the SN cumulative distribution function is generated by applying a strictly decreasing involution to the survivor function of the N distribution (Killeen & Taylor, 2004). An involution is a transformation that is its own inverse. So, for example, if T is an involution then $T\{T[1 - F(x)]\} = 1 - F(x)$.

Theorem 6.2. *SDT predicts that the area under the ROC curve (AUC) equals the probability correct in a two-sample, two-alternative forced-choice task.*

Proof. If we let $w = P(\text{FA})$ and define the function g such that $g(w) = P(\text{H})$ then

$$\begin{aligned} AUC &= \int_0^1 g(w)dw \\ &= \int_0^1 [1 - F_{\text{SN}}(X_{\text{C}})]\, d\,[1 - F_{\text{N}}(X_{\text{C}})] \\ &= \int_{+\infty}^{-\infty} [1 - F_{\text{SN}}(X_{\text{C}})] \frac{d\,[1 - F_{\text{N}}(X_{\text{C}})]}{dX_{\text{C}}} dX_{\text{C}} \qquad (6.15) \\ &= \int_{+\infty}^{-\infty} [1 - F_{\text{SN}}(X_{\text{C}})] \left[-f_N(X_{\text{C}})\right] dX_{\text{C}} \\ &= \int_{-\infty}^{+\infty} f_N(X_{\text{C}})\, [1 - F_{\text{SN}}(X_{\text{C}})]\, dX_{\text{C}} \qquad (6.16) \\ &= P(X_{\text{SN}} > X_{\text{N}}). \end{aligned}$$

The limits in Equation (6.15) are from $+\infty$ to $-\infty$ because $P(\text{FA}) = 0$ when $X_{\text{C}} = +\infty$ and $P(\text{FA}) = 1$ when $X_{\text{C}} = -\infty$. The last equality holds because the integrand in Equation (6.16) gives the likelihood that the sample from the N distribution equals X_{C} and the sample from the SN distribution is greater than this value. □

AUC is a widely used measure of bias-free classifier performance. For example, compared to d', it has a number of distinct advantages. Perhaps the most important is that AUC is a nonparametric measure that makes no assumptions about the underlying N and SN distributions. In contrast, d' is unambiguously defined only when the N and SN distributions have variances that are equal.

The two-sample, two-alternative, forced-choice task is closely related to multiple-look experiments, in which the observer is presented with r independent samples of either N or SN on each trial (e.g., Green & Swets, 1966). As in the YES–NO detection task, the observer's task is to respond YES or NO, depending on whether the r samples were all SNs or Ns. Another well-known result relates the performance of an ideal observer in the multiple-look experiment to performance in the YES–NO detection task.

Theorem 6.3. *Suppose an ideal observer with perfect memory participates in a multiple-look experiment in which r independent samples of N or SN are presented on each trial. Denote the d' of this observer in the YES–NO detection task as d'_{YN} and the d' in the multiple-look experiment as d'_r. Then the normal, equal-variance model predicts that*

$$d'_r = \sqrt{r}\, d'_{YN}. \qquad (6.17)$$

Proof. In the multiple-look experiment, each of the r N or SN samples generates its own sensory value. Denote the ith of these by x_i, and the collection of all r

by the vector $\underline{x}' = [x_1, x_2, \ldots, x_r]$. Under the assumptions of the proposition, note that on N trials, $\underline{x}$ has an r-dimensional multivariate Z distribution, and on SN trials it has an r-dimensional multivariate normal distribution with mean vector $\underline{\mu}' = [d'_{\mathrm{YN}}, d'_{\mathrm{YN}}, \ldots, d'_{\mathrm{YN}}]$ and variance–covariance matrix equal to the identity. Since the variance equals 1 in all directions, the standardized distance between the N and SN means is

$$\begin{aligned} d'_r &= \sqrt{(d'_{\mathrm{YN}} - 0)^2 + (d'_{\mathrm{YN}} - 0)^2 + \cdots + (d'_{\mathrm{YN}} - 0)^2} \\ &= \sqrt{r\, d'^2_{\mathrm{YN}}} \\ &= \sqrt{r}\, d'_{\mathrm{YN}}. \end{aligned}$$

□

Estimation of d'_r for human observers shows that it increases with r, but more slowly than predicted by Equation (6.17) (Green & Swets, 1966). The most likely reason is that human observers do not have perfect memory, and thus are unable to take full advantage of all r stimulus samples.

6.3.3 Extensions

Marr (1982) famously proposed the hierarchical classification of mathematical models as computational, algorithmic, or implementational. In mathematical psychology, Marr's algorithmic-level models are often referred to as process models. SDT provides a computational-level description of decision making, since it makes no attempt to describe the underlying algorithms or perceptual or cognitive processes that mediate decision-making. During the 1970s, great efforts were devoted to developing process models of decision-making, and currently there are several different process interpretations of SDT. Perhaps the most popular is provided by the drift-diffusion model (Link & Heath, 1975; Ratcliff, 1978), which is illustrated in Figure 6.5. The idea is that instead of representing the sensory effects of the stimulus on each trial with a single random sample from the N or SN distributions, as in classical SDT, the observer is assumed to repeatedly sample the presented stimulus as long as it is available. Each sample $\mathbf{X}$ is compared to the criterion X_{C} by computing the difference $\mathbf{X} - X_{\mathrm{C}}$, and these differences are accumulated. The sampling and accumulating processes continue until the resulting sum (or integral) first exceeds an upper criterion A or falls below a lower criterion −B (i.e., see Figure 6.5b). Sampling terminates with a YES response in the former case, and with a NO response in the latter case.

This version of the drift-diffusion model includes the d' and X_{C} parameters of SDT plus the response criteria A and B. However, in addition to predicting accuracy data, the diffusion model also predicts RTs because closed-form expressions exist for first-passage times (i.e., time when the process first crosses a response threshold). As a result, there are more data to fit, and therefore more degrees of freedom available for parameter estimation. Several computer packages are

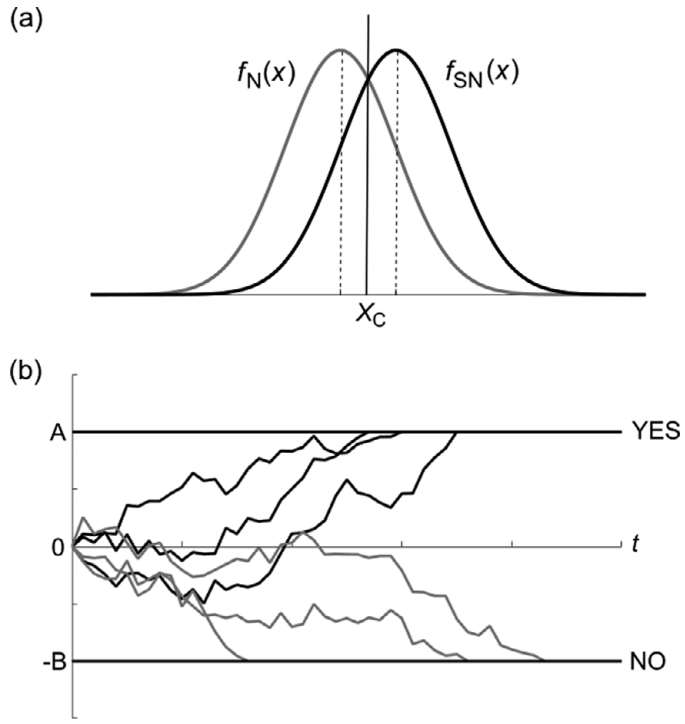


Figure 6.5 *(a) The normal, equal-variance model in which $d' = 1$. (b) A drift-diffusion model in which the drift is determined by random sampling from the N or SN distribution. Samples larger than X_C push the drift up, whereas samples smaller than X_C push it down. Sample paths are shown for six hypothetical trials – three SN trials (in black) and three N trials (in gray).*

available that automate this parameter estimation process (Vandekerckhove & Tuerlinckx, 2007; Wiecki, Sofer, & Frank, 2013).

Note that the drift-diffusion model can represent a response bias in two different ways. One is to place X_C at some point where the likelihood ratio is different from 1 (assuming equal base rates and payoffs), and another is to set A $\neq$ B. Of course, the classical SDT model can account for bias only by adjusting X_C. Consider a condition in which the observer adopts a conservative criterion and therefore is biased towards responding NO. Thus, according to SDT, X_C is set at some point where the likelihood ratio is greater than 1 (i.e., $\beta > 1$). Now consider trials in that condition where the sensory value falls at some point where the likelihood ratio is greater than 1 but less than β. According to SDT, the observer will respond NO on this trial, even though the evidence objectively favors a YES response (because the likelihood ratio is greater than 1). Balakrishnan (1999) presented evidence against this prediction. In particular, he described results of several experiments that suggested that observers always respond with the alternative that is most likely to be correct, even if they are biased towards one response and against the other.

Unfortunately, there is no way to represent this state of affairs in classical SDT. In contrast, the drift-diffusion model offers an elegant resolution to this apparent paradox. Balakrishnan's results suggest that X_C is set at the point for which $\beta = 1$ in all applications (e.g., as in Figure 6.5). A bias towards a NO response can then be implemented by setting A > B. Thus, according to this account, the evidence is always judged objectively. Evidence that objectively favors SN always makes a YES response more likely and evidence that favors N always makes a NO response more likely. Therefore, a bias towards responding NO does not color the observer's view of the world. Instead, the observer is simply willing to stop and respond NO on the basis of less overall evidence than they are willing to stop and respond YES. This more reasonable view of response bias is among the greatest advantages that the drift-diffusion model provides over and above classical SDT.

6.4 Two or More Dimensions: General Recognition Theory

SDT is useful for understanding behavior in any task in which the observer's decision is based on a single sensory dimension. Most real-world stimuli vary on multiple dimensions, however, and many perceptual decisions require attention to more than one dimension. For example, there is no single sensory dimension that allows accurate face identification. For this reason, there is obvious value in extending SDT to multiple stimulus dimensions.

At first glance, this seems like a straightforward exercise. An obvious place to begin is by replacing the unidimensional probability distributions that are used to represent the sensory effects of a stimulus in SDT with multivariate probability distributions. But complications quickly arise even in the case of two sensory dimensions. First, some sensory dimensions interact, and the perceptual literature includes a bewildering number of terms that have been proposed to describe these interactions, including perceptual independence, separability, integrality, holism, configurality, sampling independence, dimensional orthogonality, and performance parity. How should these different types of sensory interactions be modeled? And how are they all related to each other? Second, how should the decision process be modeled? In SDT, the sensory space is a line, and in two-alternative tasks, the observer is typically assumed to divide the line into two regions – one associated with each response alternative. Fortunately, there are only a few ways to do this. In fact, a standard lecture in courses on SDT is to show that almost any decision strategy is equivalent to the Equation (6.1) decision rule. However, if there are two sensory dimensions, then the sensory space is a plane, and there are an infinite number of qualitatively different ways to divide a plane into two regions.

Not surprisingly, the first attempt to generalize SDT to multiple stimulus dimensions, by Tanner in 1956, ignored most of these issues. Specifically, Tanner (1956) allowed for only one simple type of perceptual interaction and he assumed that observers always use an optimal decision rule. Despite these simplifying

assumptions, Tanner's contribution was significant because he was the first to consider multiple sensory dimensions. Even so, it was another 30 years before a more useful multidimensional version of SDT was developed. During the late 1980s, a flurry of articles significantly generalized Tanner's approach. The title of Tanner's (1956) article was "Theory of recognition." To pay homage to his contributions, Ashby and Townsend (1986) called their more general approach, general recognition theory (GRT). GRT quickly developed: Ashby and Townsend (1986) proposed a GRT-based theory of perceptual interactions, Ashby and Gott (1988) studied decision rules in multidimensional perceptual spaces, and Ashby and Perrin (1988) used GRT to develop a unified theory of similarity and identification.

6.4.1 Identification versus Categorization

GRT has been applied to a wide variety of tasks. But two tasks – identification and categorization – have emerged as the most popular, and which one is used depends on the goals of the research. In particular, identification tasks are used if the primary goal is to study perceptual representations, whereas categorization tasks are used if the primary goal is to study decision processes.

In identification tasks, there are M stimuli and M unique identifying responses. On each trial, one of the stimuli is presented, and the observer's task is to identify the stimulus by emitting the appropriate response. The data are collected in an $M \times M$ confusion matrix, in which the entry in row i and column j is the frequency with which the observer gave response j on trials when stimulus i was presented. Because the number of stimulus presentations is known, there is one constraint on each row of the confusion matrix. As a result, every confusion matrix has $M \times (M - 1)$ degrees of freedom. Note that the YES–NO detection task is a special case of this identification task in which $M = 2$ and the two stimuli to be identified are N and SN.

The most useful information in identification tasks is in the confusions that observers make, so experimental conditions are selected to guarantee errors. This is usually accomplished by using highly similar stimuli, but sometimes brief exposure durations or noise masks are used instead. Anytime one stimulus is confused for another, an error occurs. Therefore, misidentifications are most commonly made because of errors in perception, rather than because of a suboptimal decision strategy. As a result, identification tasks are a good choice if the goal is to study perceptual representations. Of course, observers can also make errors if they fail to remember which response button is associated with which stimulus. Therefore, feedback is usually provided to help observers learn these associations, and some training trials are included that are excluded from the data analysis.

In the most widely used identification tasks, the stimuli are constructed by factorially combining a small number of discrete values on two sensory dimensions. The most common choice is to factorially combine two values on two dimensions to create a total of four stimuli. Each confusion matrix collected from such a 2×2 factorial design includes 12 degrees of freedom (4×3) for parameter estimation

and model testing. If we call the two stimulus dimensions A and B, then we can denote the stimulus in which dimension or component A is at level i and component B is at level j by A_iB_j, and the corresponding response by a_ib_j.

Categorization experiments are identical to identification experiments, except they include fewer response alternatives than stimuli. In a categorization experiment, one of N stimuli is presented on each trial and the observer's task is to assign it to one of M categories, where $M < N$. The confusion matrix is therefore $N \times M$, and it contains $N \times (M-1)$ degrees of freedom. The most common choice is $M = 2$. Note that in this case, the data include N degrees of freedom. In most cases the categories are novel, in the sense that they were created specifically to use in the experiment. As a result, accurate responding requires the observer to learn the structure of these categories, most commonly via trial-by-trial feedback provided by the experimenter. Errors are most likely to occur because the observer is using a suboptimal strategy to assign stimuli to categories. Misperceptions are just as likely as in identification experiments, but they tend to have little effect on accuracy. For example, confusing one stimulus with another in the same category does not change the response, and therefore has no observable effect on behavior. For these reasons, categorization experiments are a good choice if the goal is to study decision processes.

6.4.2 Modeling Perceptual and Decisional Interactions

One of the foundational motivations for the generalization of SDT to multiple dimensions was to model perceptual interactions in a theoretically rigorous way (Ashby & Townsend, 1986). For much of the middle portion of the twentieth century, this issue was addressed almost completely in terms of operational definitions (e.g., Garner & Felfoldy, 1970; Garner, Hake, & Eriksen, 1956; Garner & Morton, 1969).[6] Ashby and Townsend (1986) created GRT principally as a theoretical structure to define perceptual independence, perceptual separability, and decisional separability. These definitions are now standard in the field. They also showed how these theoretical primitives relate to a variety of other independence-related terms that were popular in the literature.

A GRT model of the 2×2 factorial identification experiment is shown in Figure 6.6. The ellipses denote the contours of equal likelihood for the four bivariate perceptual distributions, where $f_{ij}(x_1, x_2)$ denotes the perceptual distribution associated with stimulus A_iB_j. Note that the marginal distributions associated with this stimulus are denoted by $g_{ij}(x_1)$ and $g_{ij}(x_2)$ for dimensions x_1 and x_2, respectively. Also shown are the decision bounds that divide the perceptual plane into four response regions.

According to GRT, stimulus components A and B satisfy *perceptual independence* in stimulus A_iB_j if and only if the perceived value of component

6 Use of the term "operational" is not to be confused here with the logic of operationism or converging operations (Bridgman, 1945; Von Der Heide *et al.*, 2018).

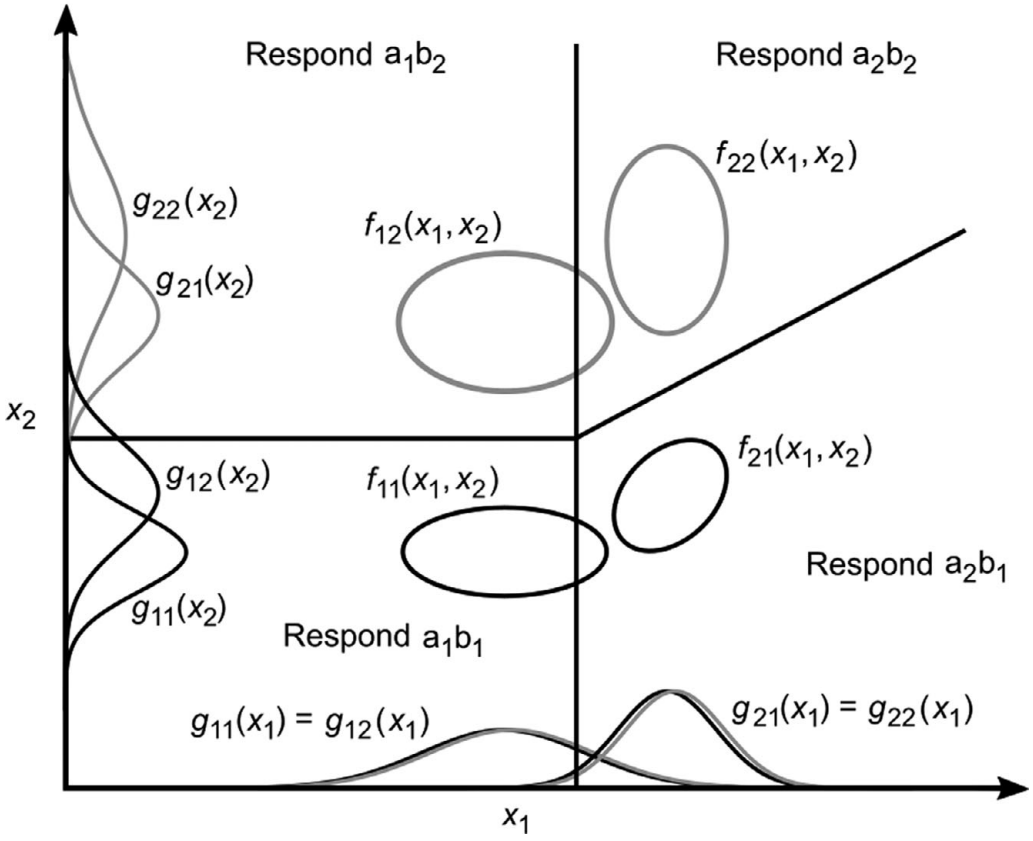


Figure 6.6 *A GRT model of the* 2×2 *factorial identification experiment. The ellipses denote the contours of equal likelihood for the four bivariate perceptual distributions.*

A is statistically independent of the perceived value of component B on trials when stimulus A_iB_j is presented. More specifically, perceptual independence of components A and B holds in stimulus A_iB_j if and only if

$$f_{ij}(x_1, x_2) = g_{ij}(x_1)g_{ij}(x_2), \tag{6.18}$$

for all values of x_1 and x_2. If perceptual independence is violated, then components A and B are perceived dependently.

Note that perceptual independence is a property of a single stimulus, in the sense, for example, that perceptual independence could hold for one stimulus and be violated for all others. In the Figure 6.6 example, the distributions are all bivariate normal, so independence is equivalent to zero correlation. Note that perceptual independence appears to be satisfied in all stimuli except A_2B_1, which displays a positive correlation between perceived values of the A and B stimulus components.

Component A is *perceptually separable* from component B if the observer's perception of A does not change when the level of B is varied. In other words, if components A and B are perceptually separable, then it is easy to attend to one and ignore the other. If this is impossible – that is, if the perception of A changes when B changes, then component A is *perceptually integral* with component B. Classic separable dimensions are color and shape, whereas classic integral dimensions are the saturation and brightness of a color patch. In GRT, all information about the perception of component A on trials when stimulus A_iB_j is presented is contained in the marginal distribution $g_{ij}(x_1)$. Therefore, component A is perceptually separable from B if and only if

$$g_{11}(x_1) = g_{12}(x_1) \text{ and } g_{21}(x_1) = g_{22}(x_1), \text{ for all values of } x_1. \tag{6.19}$$

Equation (6.19) guarantees that the perception of component A_1 is the same regardless of whether it appears with B_1 or B_2, and that the same invariance holds for component A_2. In the Figure 6.6 example, note that component A is perceptually separable from component B, but component B is not perceptually separable from component A. In particular, changing the level of B does not change the perception of A, but increasing the level of A from A_1 to A_2 increases the perceived value of component B. Note that unlike perceptual independence, perceptual separability is a property of multiple stimuli (i.e., all that share a common value on one stimulus dimension).

Finally, *decisional separability* holds on dimension x_1 if the decision about whether component A is at level 1 or level 2 does not depend on the perceived value of component B. Mathematically, this condition holds if and only if the observer uses the following decision rule to determine the level of component A:

$$\text{The level of component A is 1 if } \mathbf{X}_1 \leq X_1; \text{ otherwise, the level is 2,} \tag{6.20}$$

for some constant criterion X_1. This decision rule is equivalent to using a decision bound on dimension x_1 that is parallel to the x_2-axis (and therefore orthogonal to the x_1-axis). In the Figure 6.6 example, note that decisional separability holds on dimension x_1, but not on dimension x_2.

GRT has also been used successfully to formalize and study the notion of holistic or configural perception or processing (e.g., see discussions in Piepers & Robbins, 2012; Richler & Gauthier, 2014). GRT was first used to model the potential perceptual and decisional interactions that constitute holistic or configural perception by O'Toole, Wenger, and Townsend (2001), and it was first applied to the holistic or configural perception of faces by Wenger and Ingvalson (2002, 2003). More recently, Townsend and Wenger (2015) used GRT to propose a set of working axioms for holistic or configural perception.

As an example of how GRT has been used to study holistic processing, consider face perception, and more specifically, the composite face effect (Young, Hellawell, & Hay, 1987), which is frequently cited as a hallmark of holistic perception (Murphy, Gray, & Cook, 2017). The composite face illusion occurs in

tasks where observers are presented with an image of a face, divided into top and bottom portions roughly at the nose. Observers are asked to identify either the top or bottom half while ignoring the other half. The top and bottom portions can be drawn from either the same or different faces, the faces can be either familiar (e.g., famous) or unfamiliar, and the two halves can be either aligned or misaligned. The composite face effect is that identification of one half is impaired when the top and bottom halves are from different faces, and this impairment is greatest when the two halves are from different familiar identities.

The first step in modeling the composite face effect with GRT is to represent the space of perceptual evidence supporting identification of the two halves. For simplicity, consider the simplest case in which the top and bottom halves are always aligned. Let component A denote the top half face and component B denote the bottom half, with the subscript denoting the identity of the face. So in stimuli A_1B_1 and A_2B_2, the top and bottom halves are from the same face, whereas in stimuli A_1B_2 and A_2B_1, the two halves are from different faces.

The next step is to construct a null model that does not display any type of holism or configurality. We do this by assuming perceptual independence for all stimuli, perceptual and decisional separability on both dimensions, and that all variances are equal. In this model, which is illustrated in Figure 6.7a, the identity of the top half of the stimulus does not affect the perceptual representation or the decision made about the bottom half.

The final step is to build a model that assumes holistic perception when the top and bottom halves are from the same face, but not when they mismatch. There are several ways to do this. One is to assume a positive perceptual dependence when the two halves match and a negative dependence when they mismatch. This model, which is illustrated in Figure 6.7b, corresponds to the type of within-stimulus relationships that are implied in the vernacular use of holism, configurality, or Gestalt (O'Toole, Wenger, & Townsend, 2001; Townsend & Wenger, 2015). A second way is to change the marginal means of the distributions such that confusability increases when the bottom and top are mismatched and decreases when they are matched (Figure 6.7c). The same effect could be obtained by the third possible way of modeling holism: by shifting the decision bounds (Figure 6.7d). Of course, these possibilities could also be combined in a variety of ways.

Two significant points have been made by applying GRT to the issue of holism. The first is that, just as there are varieties of independence in perception (Ashby & Townsend, 1986), there are a variety of ways to obtain patterns of data from which one can infer holism or configurality. The second is that analysis of a task by way of GRT can provide important insights into the extent to which the task is capable of testing a hypothesis. For example, GRT simulations reported by Richler *et al.* (2008) demonstrated that the standard method of testing the composite face effect (see discussions in Richler & Gauthier, 2013; Rossion, 2013) does not provide data that would allow for testing the strong hypothesis that holism is a within-stimulus effect.

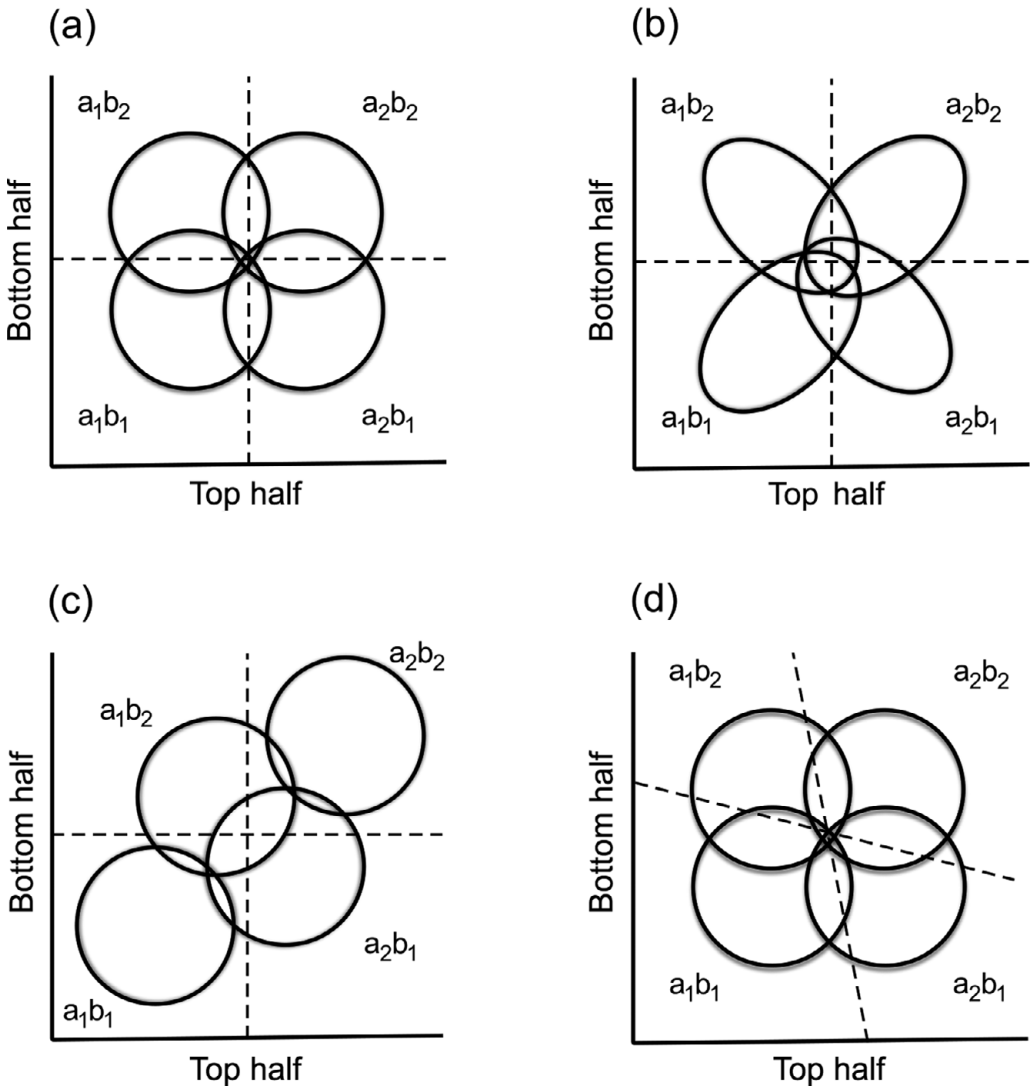


Figure 6.7 *Alternative GRT models of the holism or configurality thought to underlie the composite face effect: (a) lack of holism or configurality; (b) positive perceptual dependencies when the halves match and negative perceptual dependencies when they mismatch; (c) shifting the perceptual means to model increased accuracy when the halves match and decreased accuracy when they mismatch; (d) accounting for increased accuracy when the halves match by shifting the decision bounds.*

6.4.3 Applications to Categorization Tasks

In principle, the application of GRT to categorization tasks is the same as its application to identification. In both cases, the data are summarized in a confusion matrix, and the primary focus is on the pattern of errors made by observers.

One important statistical difference however, is that, for the same number of stimuli, categorization data have fewer degrees of freedom – often far fewer. For example, the most common categorization experiments include two categories. Therefore, with M stimuli, an identification confusion matrix includes $M \times (M-1)$ degrees of freedom and the corresponding categorization confusion matrix includes only M degrees of freedom (i.e., since the confusion matrix has order $M \times 2$). Because the data include fewer degrees of freedom, GRT applications to categorization tasks include simplifying assumptions that reduce the number of free parameters, relative to GRT applications to identification data.

In fact, when applied to categorization data, the most common assumption is that all perceptual representations are multivariate normally distributed with known means and with variance–covariance matrix equal to σ^2I, where σ^2 is the common noise variance on each dimension and I is the identity matrix. Thus, only one free parameter is typically assigned to model all perceptual representations (i.e., σ^2), and all other parameters are used to model decision bounds. This choice reflects the assumption that in categorization experiments, errors are more likely caused by suboptimal decision strategies than by faulty perception. Allocating the lion's share of parameters to the decision bounds provides the best opportunity to characterize these suboptimalities.

The mean of each perceptual distribution describes the mean perceived value of each stimulus. In some cases, these could come from previous multidimensional scaling or psychophysical modeling of the stimuli. For example, in the case of sine-wave gratings (such as Gabor patches) that vary in spatial frequency and orientation, a psychophysical model that describes the transformation from stimulus space to perceptual space was provided by Treutwein, Rentschler, and Caelli (1989). Another possibility, especially for dimensions that are perceptually separable, is to use Stevens' exponent. For example, the Stevens exponent for brightness is 0.33, so the mean brightness of each stimulus could be computed from $kI^{0.33}$, where I is the physical intensity of the stimulus and k is an arbitrary constant that can be set for convenience. When GRT models are fit to categorization data under these assumptions about the perceptual representations, they are often referred to as decision bound models. One advantage they have over GRT models with more complex perceptual representations, which is illustrated in the next result (due to Ashby & Maddox, 1993), is that no numerical integration is needed to fit any of the most common models.

Theorem 6.4. *Consider a categorization task with two categories, A and B, and a decision-bound model with one linear boundary. Let the random vector* $\underline{\boldsymbol{X}}_i$ *denote the perceived value of stimulus* S_i*. Assume that* $\underline{\boldsymbol{X}}_i$ *has a multivariate normal distribution with known mean* $\underline{\mu}_i$ *and variance–covariance matrix* $\sigma^2 I$*. Then the decision bound is the set of all points that satisfy*

$$h(\mathbf{X}_i) = \underline{b}'\mathbf{X}_i + c = 0, \tag{6.21}$$

for some vector of constants $\underline{b}$ *and constant c. This model, called the general linear classifier, predicts that*

$$P(A|S_i) = \Phi\left(\frac{\underline{b}'\underline{\mu}_i + c}{\sigma\sqrt{\underline{b}'\underline{b}}}\right), \tag{6.22}$$

where Φ *is the cumulative distribution function of a standard normal (i.e., Z) distribution.*

Proof. Under the conditions of the proposition, the decision rule of the general linear classifier is "Respond A if $h(\underline{\mathbf{X}}_i) > 0$; otherwise, respond B." Therefore

$$P(A|\mathrm{S}_i) = P\left[h(\underline{\mathbf{X}}_i) > 0|\mathrm{S}_i\right]. \tag{6.23}$$

Now $\underline{\mathbf{X}}_i$ has a multivariate normal distribution with mean vector $\underline{\mu}_i$ and variance–covariance matrix $\sigma^2\mathrm{I}$. As a result, $h(\underline{\mathbf{X}}_i)$ has a univariate normal distribution with mean $\underline{\mathrm{b}}'\underline{\mu}_i + c$ and variance $\sigma^2\underline{\mathrm{b}}'\underline{\mathrm{b}}$. The result follows immediately from these observations. □

Since the $\underline{\mu}_i$ are assumed to be known, the parameters of the model are the noise variance σ^2 and the decision-bound parameters $\underline{\mathrm{b}}$ and c. If there are r perceptual dimensions, then $\underline{\mathrm{b}}$ has order $r \times 1$. However, without loss of generality, one entry in $\underline{\mathrm{b}}$ can be set arbitrarily, so $\underline{\mathrm{b}}$ has only $r - 1$ free parameters.[7] Therefore, if the perceptual space is two-dimensional, this model has three free parameters (i.e., one slope parameter, the decision-bound intercept c, and the noise variance σ^2).

Predictions for the decision-bound model that assumes a quadratic decision bound, called the general quadratic classifier, were derived by Ashby and Maddox (1993). Predictions for models that assume some form of decisional separability can be found in Ashby and Valentin (2018). For these models, the decision bound is compatible with an explicit rule that is easily verbalized. For example, the rule: "Respond A if $\mathbf{X}_1 > c_1$ and $\mathbf{X}_2 > c_2$; otherwise, respond B" is equivalent to the conjunction rule "Respond A if the stimulus is large on both dimensions; otherwise, respond B." Ashby and Valentin (2018) also described predictions of models that assume the participant guesses randomly on every trial.

Criterial noise can be added to decision-bound models by assuming that the decision rule is "Respond A if $h(\underline{\mathbf{X}}) > \boldsymbol{\epsilon}_c$; otherwise, respond B," where $\boldsymbol{\epsilon}_c$ is normally distributed with mean 0 and variances σ_c^2. If the decision bound is linear, then it is straightforward to show that perceptual and criterial noise are not separately identifiable (Ashby & Maddox, 1993). Instead, only the sum of the perceptual and criterial noise variances is estimable. For this reason, it makes no difference whether we assume that the noise is perceptual or decisional (or some combination of the two). Once predicted probabilities are computed, the

7 For example, assume $r = 2$. Then note that at least one of b_1 and b_2 (i.e., the entries in $\underline{\mathrm{b}}$) must be nonzero. Note that the decision rule "Respond A if $h(\underline{\mathrm{X}}) > 0$" is unchanged if we divide both sides by a positive constant. Therefore, without loss of generality, we can divide both sides by $\sqrt{b_1^2 + b_2^2}$. Note that the sum of the squared entries in the revised $\underline{\mathrm{b}}$ vector now equals 1. As a result, we can always replace b_2 with $\sqrt{1 - b_1^2}$.

parameters can be estimated by finding the numerical values that maximize the likelihood-related statistic L^* in Equation (6.33) below.

6.4.4 Applications to Identification Tasks

GRT has been used to analyze data from identification confusion matrices in two different ways. One approach is to compute certain summary statistics from the empirical confusion matrix and then to check whether these satisfy conditions that are characteristic of perceptual independence, perceptual separability, or decisional separability. The other approach is to fit GRT models to the entire confusion matrix. To test various assumptions about perceptual and decisional processing – for example, to test whether perceptual independence holds – a version of the model that assumes perceptual independence is fit to the data as well as a version that makes no assumptions about independence. This latter version contains the former version as a special case (i.e., in which all covariance parameters are set to zero), so it can never fit worse. After fitting these two models, we conclude that perceptual independence is violated if the more general model fits significantly better than the more restricted model that assumes perceptual independence (Ashby & Perrin, 1988; Thomas, 2001).[8] Because these approaches are so different, we discuss each in turn.

It is important to note, however, that regardless of which method is used, there are certain non-identifiabilities in GRT models that could limit the conclusions that are possible to draw from any such analyses (e.g., Menneer, Wenger, & Blaha, 2010; Silbert & Thomas, 2013). The problems are most severe when GRT is applied to identification data from 2×2 factorial designs (i.e., with stimuli A_1B_1, A_1B_2, A_2B_1, and A_2B_2). For example, Silbert and Thomas (2013) showed that in 2×2 applications where there are two intersecting linear decision bounds that do not satisfy decisional separability, there always exists an alternative model that makes the exact same empirical predictions and satisfies decisional separability (and these two models are related by a linear transformation). Thus, in standard applications of GRT to identification experiments that use a 2×2 factorial design, decisional separability is not testable, nor are the slopes of the decision bounds uniquely estimable. It turns out, however, that for a variety of reasons, these non-identifiabilities are not catastrophic.

First, there are no identifiability problems if the perceptual dimensions are known. Obviously, the linear transformation that rotates intersecting linear bounds so that one is vertical and one is horizontal also rotates the perceptual dimensions. So although decisional separability holds in the new model, the separability is with respect to novel dimensions. In other words, one interpretation of the identifiability problem is that if the best-fitting GRT model to some single

8 Note that many of the statistical packages written for estimating GRT models provide estimates of parameter variability and/or confidence intervals, allowing one to determine whether (for example) a parameter estimate can be inferred to be reliably different from 0.

confusion matrix collected in an experiment that used a 2×2 factorial design assumes intersecting linear bounds that violate decisional separability, then there is always an alternative GRT model that fits equally well and assumes that the observer made decisions by selectively attending to some different perceptual dimensions. With complex stimuli, such as faces, this will often be difficult to rule out. However, with many simple stimuli, this possibility is straightforward to reject. For example, consider sine-wave gratings (e.g., such as Gabor patches) that are created by factorially combining two spatial frequencies (bar widths) and two (bar) orientations. An enormous visual perception literature tells us that humans treat these two dimensions as primary (e.g., DeValois & De Valois, 1990). So any conclusions about decisional separability drawn from a GRT analysis should be immune to identifiability problems because the mathematically equivalent model that makes different assumptions about decisional separability must assume that the observer perceived the stimuli in a way that is incompatible with the visual perception literature.

Second, the problems do not generally exist with 3×3 or larger factorial designs (as used e.g., by Ashby *et al.*, 2001). In the 3×3 case, the GRT model with linear bounds requires at least four decision bounds to divide the perceptual space into nine response regions (e.g., in a tic-tac-toe configuration). Typically, two will have a generally vertical orientation in the two-dimensional perceptual space and two will have a generally horizontal orientation. Linear transformations will rotate the vertical-tending bounds by the same amount, and the horizontal-tending bounds by the same amount. Therefore, unless the two vertical-tending bounds are parallel and the two horizontal-tending bounds are parallel, there is no linear transformation that guarantees decisional separability for all four bounds. For example, if the two vertical-tending bounds are not parallel, then the linear transformation that makes one perfectly vertical (guaranteeing decisional separability) will leave the other oblique to the abscissa (causing a violation of decisional separability). Thus, in 3×3 (or higher) designs, decisional separability is typically identifiable and testable.

Third, there are simple experimental manipulations that can be added to the basic 2×2 identification experiment to test for decisional separability. Currently, more than 30 different qualitative differences have been identified in the learning and performance of tasks in which observers use strategies that satisfy versus violate decisional separability (for a review of most of these, see Ashby & Valentin, 2017). For example, switching the locations of the response buttons interferes with performance if decisional separability fails more than if decisional separability holds (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004), and delaying feedback by a few seconds has a similar effect, but on learning, rather than performance (Crossley & Ashby, 2015; Dunn, Newell, & Kalish, 2012; Maddox, Ashby, & Bohil, 2003; Maddox & David, 2005).

Fourth, one could analyze the 2×2 data using GRT-wIND (GRT with INDividual differences; Soto *et al.*, 2015), which was inspired by the INDSCAL model of

multidimensional scaling (Carroll & Chang, 1970). Like INDSCAL, GRT-wIND is fit to the data from all individuals simultaneously. All observers are assumed to share the same group perceptual distributions (see Silbert & Thomas, 2017 for a discussion of this assumption), but different observers are allowed different linear bounds and they are assumed to allocate different amounts of attention to each perceptual dimension. The model does not suffer from the identifiability problems identified by Silbert and Thomas (2013), even in the 2×2 case, because with different linear bounds for each observer, there is no linear transformation that simultaneously makes all these bounds satisfy decisional separability.

Summary Statistics Approach

The first approach that used GRT to test perceptual and decisional assumptions was based on parametric and nonparametric summary statistics that were derived from the identification–confusion matrix (see, e.g., Figure 11, p. 172 of Ashby & Townsend, 1986). This later evolved to an approach known as multidimensional signal detection analysis (MSDA; Kadlec, 1995; Kadlec & Townsend, 1992a, 1992b), which extended the concepts originally presented by Ashby and Townsend (1986) and combined those equalities with tests of equalities on Gaussian SDT parameters. This was later both simplified and refined, as summarized by Silbert and Hawkins (2016), under the strong assumption that decisional separability always holds (see also Silbert & Thomas, 2013).

The most popular summary statistics tests use measures called *marginal response invariance* and *report independence* to draw inferences about perceptual separability and perceptual independence. Marginal response invariance holds at the *i*th level of the first dimension if the following equality holds:

$$\begin{aligned} P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_1) &= P(\mathrm{a}_i\mathrm{b}_1|\mathrm{A}_i\mathrm{B}_1) + P(\mathrm{a}_i\mathrm{b}_2|\mathrm{A}_i\mathrm{B}_1) \\ &= P(\mathrm{a}_i\mathrm{b}_1|\mathrm{A}_i\mathrm{B}_2) + P(\mathrm{a}_i\mathrm{b}_2|\mathrm{A}_i\mathrm{B}_2) \\ &= P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_2), \end{aligned} \tag{6.24}$$

where, as before, $P(\mathrm{a}_k\mathrm{b}_m|\mathrm{A}_i\mathrm{B}_j)$ is the probability that the participant responded $\mathrm{a}_k\mathrm{b}_m$ on trials when stimulus $\mathrm{A}_i\mathrm{B}_j$ was presented. Marginal response invariance provides information about perceptual separability so long as decisional separability holds. If decisional separability does hold, then a failure of marginal response invariance at any level of a given dimension implies that perceptual separability fails on that dimension (Ashby & Townsend, 1986). If the marginal d's are also unequal on that dimension, then our conclusion that perceptual separability fails is further bolstered.

Before GRT, the most popular method for assessing separability was via a categorization task called the *filtering task*, which uses the same stimuli as the 2×2 identification task, but asks observers to report the level of component A or the level of component B, rather than identifying the stimulus uniquely. Ashby and Maddox (1994) proposed an RT version of marginal response invariance for this

task that they called *marginal RT invariance*. Specifically, for $i = 1$ or 2, marginal RT invariance holds for component A if

$$P(\mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1, \mathrm{a}_i) = P(\mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2, \mathrm{a}_i), \text{ for all } t > 0, \tag{6.25}$$

where **RT** is the RT and a_i indicates that the observer responded that the level of component A was i. Ashby and Maddox (1994) showed that if decisional separability holds and if RT decreases with the distance from the percept to the decision bound – an assumption called the RT–distance hypothesis – then perceptual separability holds if and only if marginal RT invariance is satisfied for both correct and incorrect responses.

Ashby and Maddox (1994) only investigated tasks with two response alternatives (i.e., the filtering and redundancy tasks popularized by Garner, 1974). Townsend, Houpt, and Silbert (2012) applied a similar approach to the 2×2 identification task. They defined an RT invariance condition similar to marginal RT invariance that they called timed marginal response invariance. This condition holds in the 2×2 identification task for level i of component A if, for all $t > 0$:

$$\begin{aligned} P(\mathrm{a}_i\mathrm{b}_1, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1) + P(\mathrm{a}_i\mathrm{b}_2, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1) \\ = P(\mathrm{a}_i\mathrm{b}_1, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2) + P(\mathrm{a}_i\mathrm{b}_2, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2). \end{aligned} \tag{6.26}$$

Rather than assume the RT–distance hypothesis, Townsend, Houpt, and Silbert (2012) investigated predictions of a general class of models that assumed processing of the two stimulus components occurs in parallel. Within this class of models, they showed that if perceptual and decisional separability hold then timed marginal response invariance must also hold.[9]

The parallel models considered by Townsend, Houpt, and Silbert (2012) are grounded on the assumptions of stochastic linear systems, in which the activation in a channel is proportional to the magnitude of its input (Townsend & Wenger, 2004; Wenger & Townsend, 2006). There is a channel for each level of every stimulus component, and each channel accumulates evidence that the relevant stimulus component is at the level to which the channel is tuned. In GRT and signal detection theory, if the likelihood ratio is monotonic, then evidence increases with distance from the boundary (or criterion). For this reason, the parallel models considered by Townsend, Houpt, and Silbert (2012) are closely related to the models that Ashby and Maddox (1994) considered, which satisfy the RT–distance hypothesis.

Given this similarity, it is not surprising that marginal RT invariance and timed marginal response invariance are closely related. First, note that

$$P(\mathrm{a}_i\mathrm{b}_1, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_j) + P(\mathrm{a}_i\mathrm{b}_2, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_j) = P(\mathrm{a}_i, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_j). \tag{6.27}$$

9 Note that this result is weaker than the if and only if result that is possible if the RT–distance hypothesis is assumed to hold in the filtering task.

Next note that marginal RT invariance is equivalent to assuming that for all $t > 0$:

$$\frac{P(\mathrm{a}_i, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1)}{P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_1)} = \frac{P(\mathrm{a}_i, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2)}{P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_2)}. \tag{6.28}$$

Now Townsend, Houpt, and Silbert (2012) showed that if timed marginal response invariance holds then so does marginal response invariance [i.e., Equation (6.24)]. Therefore, if the joint probabilities in the numerators of Equation (6.28) are equal for all t, then the probabilities in the denominators are also equal. Therefore, when applied to the filtering task, marginal RT invariance and timed marginal response invariance are equivalent.

The summary statistics described so far are targeted at testing for perceptual separability. Another set of statistics targets perceptual independence. Report independence (called sampling independence in the early literature) is assessed for each individual stimulus and provides information about perceptual independence, again assuming that decisional separability holds. Report independence holds in the 2×2 identification task for stimulus $\mathrm{A}_i\mathrm{B}_j$ if:

$$\begin{aligned} P(\mathrm{a}_i\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j) &= P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_j) \times P(\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j) \\ &= \left[P(\mathrm{a}_i\mathrm{b}_1|\mathrm{A}_i\mathrm{B}_j) + P(\mathrm{a}_i\mathrm{b}_2|\mathrm{A}_i\mathrm{B}_j)\right] \\ &\quad \times \left[P(\mathrm{a}_1\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j) + P(\mathrm{a}_2\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j)\right]. \end{aligned} \tag{6.29}$$

Ashby and Townsend (1986) showed that if decisional separability holds, then a failure of report independence implies a violation of perceptual independence.

Townsend, Houpt, and Silbert (2012) also proposed an RT-invariance condition that is similar to report independence. Specifically, timed report independence holds for the response $\mathrm{a}_i\mathrm{b}_j$ with stimulus $\mathrm{A}_k\mathrm{B}_m$ if for all times $t > 0$:

$$\begin{aligned} &P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m, \mathrm{a}_i\mathrm{b}_j) \times P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m) \\ &\quad = P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m, \mathrm{a}_i) \times P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m, \mathrm{b}_j). \end{aligned} \tag{6.30}$$

Townsend, Houpt, and Silbert (2012) showed that, within the class of parallel models they were considering, if decisional separability and perceptual independence both hold then timed report independence must hold.

Summary statistics approaches are complemented by the model-fitting approach described next. Indeed, since at least the work of Thomas (2001), there has been a focus on combining summary statistics and Gaussian model-fitting as complementary sources of converging evidence in supporting inferences (Cornes *et al.*, 2011; Von Der Heide *et al.*, 2018; Wenger & Rhoten, 2020).

Fitting the Gaussian Model to Identification Data

As mentioned earlier, a second approach for analyzing data from identification experiments is to fit a variety of different GRT models to the confusion matrices.

Assumptions about perceptual interactions and decision processes can be tested by comparing model fits of nested models in which the restricted model makes some specific assumption, such as perceptual separability, and the more general model does not (Ashby & Perrin, 1988; Thomas, 2001). The primary advantage of this approach over the summary statistics approach is that, although it is parametric, it makes fewer assumptions about perceptual and decisional processes, and therefore should be less prone to false conclusions. The trade-off though is that it is more computationally intensive, since it often requires numerical integration.

This model-fitting approach is necessarily parametric since numerical predictions are possible only if a specific functional form is specified for the perceptual distributions. All previous applications of this approach have assumed that the perceptual distributions are multivariate normal. Furthermore, all applications have assumed that there are only two relevant sensory dimensions. In principle, the fitting algorithms (described below) are straightforward to extend to more than two dimensions, but such models could include many more free parameters and therefore significantly increased computation time. Thus, to date, applications that have fit GRT models to identification confusion matrices have assumed that the sensory distributions are bivariate normal pdfs, and the response regions are defined on a plane. A variety of different assumptions about decision processes are possible. Figure 6.8 illustrates six of these.

In an identification experiment with M stimuli, the resulting confusion matrix includes $M \times (M-1)$ degrees of freedom. As a result, this value fixes the maximum number of parameters that can be estimated. A bivariate normal distribution has a maximum of five free parameters – a mean and variance on both dimensions, and a covariance. Therefore, the smallest value of M for which the most general possible GRT model can be estimated is $M = 6$. In this case, the 6×6 confusion matrix has 30 degrees of freedom, and the six perceptual distributions needed to model the perceptual effects of the six stimuli have 30 parameters. However, the origin and unit of measurement on each perceptual dimension are arbitrary. Therefore, without loss of generality, the means on both dimensions can be set to 0 in any one perceptual distribution (to set the origin) and the variances in that distribution can be set to 1 (to set the unit of measurement). This reduces the number of free parameters to 26 (i.e., to $5M - 4$), which leaves a maximum of four parameters to model the decision-bounds and assess the validity of the model. The fact that only four degrees of freedom remain rules out some decision models (e.g., the general quadratic classifier), but not all. For example, in Figure 6.8, the minimum distance and optimal classifiers have no free decision parameters, and the model that assumes decisional separability has only two free parameters (i.e., the two intercepts).

As the order of the confusion matrix increases above six, the degrees of freedom increases faster than the number of free parameters in the full GRT model. As a result, the larger the matrix, the more extra degrees of freedom there are to estimate decision-bound parameters and to test the validity of the model. For example, Ashby *et al.* (2001) fit the full model to a variety of different 9×9 confusion

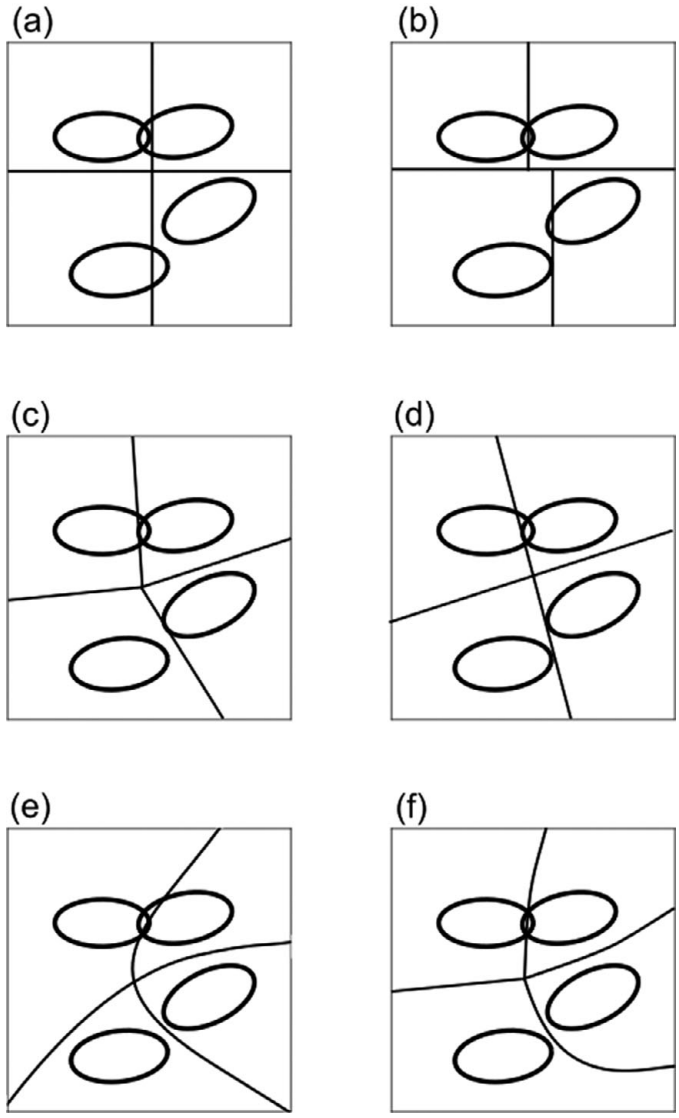


Figure 6.8 *Different types of decision bounds used in GRT modeling. (a) Decisional separability is satisfied on both dimensions. (b) Decisional separability is satisfied on dimension 2, but not on dimension 1. (c) Decision bounds of the minimum distance classifier. (d) Decision bounds of the general linear classifier. (e) Decision bounds of the general quadratic classifier. (f) Decision bounds of the optimal classifier.*

matrices, which each have 72 degrees of freedom, and with nine stimuli the full model has only 41 free perceptual parameters.

On the other hand, note that the 2×2 factorial design, which as previously mentioned is the most popular identification experiment, includes only four stimuli.

Therefore, the full model includes 16 free perceptual parameters (i.e., $5 \times 4 - 4$) and each confusion matrix includes only 12 degrees of freedom (i.e., 4×3). As a result, the full GRT model is not estimable in these experiments. So when GRT models are fit to single confusion matrices from 2×2 factorial designs, some assumptions must be made to reduce the number of free parameters.

When fitting any GRT model to identification data, parameter estimation is accomplished via the method of maximum likelihood. Denote the M stimuli by $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_M$ and the corresponding M responses by $\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_M$. Let n_{ij} denote the entry in row i and column j of the confusion matrix – that is, the frequency with which the observer responded R_j on trials when stimulus S_i was presented. Note that the n_{ij} are random variables, and the entries in each row of the confusion matrix have a multinomial distribution. In particular, if $P(\mathrm{R}_j|\mathrm{S}_i)$ is the true probability that response R_j is given on trials when stimulus S_i is presented, then the probability of observing the response frequencies $n_{i1}, n_{i2}, \ldots, n_{iM}$ in row i is

$$P(n_{i1}, n_{i2}, \ldots, n_{iM}|\mathrm{S}_i) = \frac{N_i!}{n_{i1}!\ n_{i2}! \cdots n_{iM}!} P(\mathrm{R}_1|\mathrm{S}_i)^{n_{i1}} P(\mathrm{R}_2|\mathrm{S}_i)^{n_{i2}} \cdots P(\mathrm{R}_M|\mathrm{S}_i)^{n_{iM}}, \tag{6.31}$$

where N_i is the total number of stimulus S_i presentations (i.e., so $N_i = \sum_j n_{ij}$). The probability or likelihood of observing the entire confusion matrix is the product of the probabilities of observing each row:

$$L = \prod_{i=1}^{M} P(n_{i1}, n_{i2}, \ldots, n_{iM}|\mathrm{S}_i). \tag{6.32}$$

In all Gaussian GRT models, $P(\mathrm{R}_j|\mathrm{S}_i)$ is computed by integrating a multivariate normal pdf over some response region, but different models make different assumptions about the pdf and about the shape of the region. The maximum likelihood parameter estimates are the numerical values of all model parameters that maximize the likelihood L of Equation (6.32).

Two simplifications are common. First, some of the $P(\mathrm{R}_j|\mathrm{S}_i)^{n_{ij}}$ could be very small numbers, so it is common to find parameter values that maximize $\log L$ rather than L. Since log is a monotonic transformation, the parameter values that maximize L will also maximize $\log L$. Second, note that the factorial terms in Equation (6.31) do not depend on the values of any model parameters, and therefore they are typically excluded from the parameter estimation process. Therefore, the common practice is to find the maximum likelihood estimates of all parameters by maximizing the monotonically related term

$$L^* = \sum_{i=1}^{M} \sum_{j=1}^{M} n_{ij} \log P(\mathrm{R}_j|\mathrm{S}_i), \tag{6.33}$$

where as already mentioned, the predicted probabilities $P(R_j|S_i)$ are computed by integrating under the multivariate normal pdf that models the sensory representation of stimulus S_i over the R_j response region.

The difficulty of computing the integrals required to maximize L^* depends on the nature of the decision bounds assumed by the model. Decisional separability simplifies things considerably because then the integral under a bivariate normal pdf reduces to the integral under a univariate normal marginal pdf. Under these conditions, Wickens (1992) derived the first and second derivatives necessary to estimate parameters of the model quickly using the Newton–Raphson method. In models that do not assume decisional separability, the integrals are under the bivariate normal pdf over irregularly shaped regions of the plane. As a result, numerical integration is required.

Ennis and Ashby (2003) proposed an efficient algorithm for evaluating these integrals that can be used to estimate the parameters of virtually any GRT model via standard minimization software. This algorithm was described in detail by Ashby and Soto (2015). Briefly, however, the algorithm, which is described in Figure 6.9, includes the following five steps.

(1) A set of D Z-values are preloaded into an array. Each Z-value is chosen to be the center of an interval that has equal area under the Z distribution (i.e., under the pdf of a normal distribution with mean 0 and variance 1). The Cartesian product of this array with itself creates a grid of points in multidimensional space that are each the center of a rectangle (or hyper-rectangle) that all have equal volumes under the multidimensional Z pdf (i.e., the gray points on the right side of Figure 6.9). If the GRT model assumes r perceptual dimensions then after this step there will be D^r grid points. For example, to fit two-dimensional GRT models, Ashby *et al.* (2001) set $D = 100$, which creates a grid of 10,000 points in bivariate Z-space, each of which is the center of a rectangle with volume 0.0001 (i.e., 0.01^2).

(2) Note that GRT assumes that all entries in each row of a confusion matrix are computed by integrating under the same perceptual distribution. Different columns are associated with different response regions. The algorithm works row-by-row. The idea is to transform the perceptual distribution associated with the current row to a multivariate Z-distribution. This can always be accomplished via an affine transformation in which the linear transformation is based on the Cholesky factorization of the distribution's variance–covariance matrix. The second step is to compute this affine transformation.

(3) Apply this affine transformation to the decision bounds. Since affine transformations preserve linearity, this step will convert linear bounds in perceptual space to linear bounds in Z-space.

(4) Step through all D^r grid points and for each one, identify its associated response region. Each bound defines a discriminant function that assigns positive values to all points on one side and negative values to all points on the other side. With multiple bounds, each response region is characterized by a unique set of

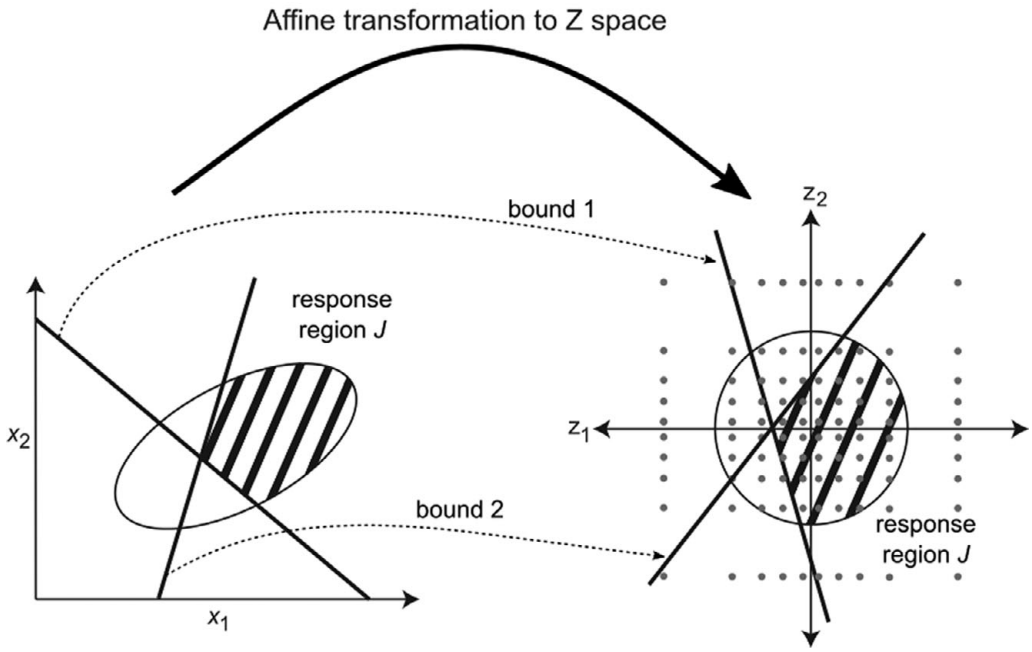


Figure 6.9 *Schematic illustration of how numerical integration is performed in the multivariate normal GRT model via Cholesky factorization.*

positive and negative discriminant values. So the response region of a point can be identified by examining its pattern of positive and negative discriminant values of all decision bounds after they have been transformed to Z-space.

(5) Suppose the current grid point is identified as belonging to response region J. The final step is to increment the integral associated with response J by $1/D^r$.

The problems caused by insufficient degrees of freedom in 2×2 factorial designs disappear if GRT-wIND (Soto *et al.*, 2015) is used instead of the traditional GRT model. GRT-wIND is fit simultaneously to the individual confusion matrices of all observers. Soto *et al.* (2017) developed an R package that fits this model using only a few commands. GRT-wIND assumes that all observers share the same group perceptual representation, which is described by the full GRT model, even in 2×2 factorial designs. Thus, GRT-wIND assumes that basic perceptual properties, such as perceptual separability and perceptual independence, or their violations, are shared by all observers. The model assumes that different observers produce different confusion matrices for two reasons – they allocate different amounts of attention to the two perceptual dimensions, and they use different decision bounds. Thus, fitting the model returns estimates of (1) the group perceptual representation (i.e., the full GRT perceptual model), (2) the total amount of attention allocated to the task by each observer, (3) the proportion of attention allocated to the two perceptual dimensions by each observer, and (4) unique decision bounds for each

observer. Soto *et al.* (2015) fit GRT-wIND to the confusion matrices of 24 different observers in a face identification experiment that used a 2 × 2 factorial design in which the four stimulus faces were created by crossing two facial identities with two emotional expressions. The 24 matrices included a total of 288 degrees of freedom (i.e., 24 × 12). The GRT-wIND model included an average of 6.67 free parameters for each individual confusion matrix, which is less than typical applications of traditional GRT models to 2 × 2 designs. GRT-wIND accounted for 99.52% of the variance in the 24 confusion matrices. Even more impressively, GRT-wIND provided a better fit than the best-fitting traditional GRT model to the data of 18 of the 24 participants.[10] Furthermore, GRT-wIND suggested that in this group of 24 observers, emotional expression was perceptually separable from facial identity, but identity was not separable from expression. In contrast, a traditional GRT analysis could only report how many of the individual participants showed this pattern.

GRT accounts of identification data have been spectacularly successful. For most of the last four decades of the twentieth century, the most successful model of identification, by far, was the Luce–Shepard choice model (Luce, 1963; Shepard, 1957), which assumes that

$$P(\mathrm{R}_j|\mathrm{S}_i) = \frac{\eta_{ij}\beta_j}{\sum_{k=1}^{M}\eta_{ik}\beta_k}, \tag{6.34}$$

where η_{ij} is the similarity between stimuli S_i and S_j and β_j is the bias toward response R_j (without loss of generality, one can set $\eta_{ii} = 1$ for all values of i and $\sum \beta_j = 1$). To ensure that the model is testable, similarity is assumed to be symmetric (i.e., so that $\eta_{ij} = \eta_{ji}$ for all values of i and j). The Luce–Shepard choice model was so successful that for many years, it was the standard against which competing models were compared. For example, in 1992, J. K. Smith summarized its performance by concluding that it "has never had a serious competitor as a model of identification data. Even when it has provided a poor model of such data, other models have done even less well" (J. K. Smith, 1992, p. 199). Even so, the model was never considered completely satisfactory – primarily because a good fit provides little insight into the psychological processes of the observer producing the data. The model merely says that the probability of confusing stimulus S_j for S_i is proportional to the product of the similarity between the two stimuli and the bias toward response R_j [the denominator in Equation (6.34) is just a normalizing constant]. Also, note that the model makes no predictions about how a decision is reached. It simply predicts the proportion of R_j responses to expect over the course of a large number of S_i trials.

GRT provided the first models that ended the dominance of the Luce–Shepard choice model, at least for identification data collected from experiments with stimuli that differed on only a couple of stimulus dimensions. In virtually every

10 This is because the full traditional-GRT model is not estimable in 2 × 2 designs, but the full GRT-wIND model is estimable.

such comparison, the GRT model provided a substantially better fit than the Luce–Shepard choice model, in many cases with fewer free parameters (Ashby *et al.*, 2001). Even so, it is important to note that the Luce–Shepard choice model is still valuable, especially in the case of identification experiments in which the stimuli vary on many unknown stimulus dimensions.

6.4.5 Extensions to Response Time

Like SDT, GRT was originally developed to account exclusively for accuracy data. Even so, there have been a number of extensions of the theory that attempt also to account for RTs. These are generally of two types. One approach is to add assumptions to GRT that allow the theory to make RT predictions but are not detailed enough to account for psychological process. Thus, like the original version of GRT (and SDT), the resulting models are descriptive, or in the language of Marr (1982), computational. The other approach is to add enough structure to GRT to model psychological process – thereby producing models that Marr identified as algorithmic. We briefly review both types in turn.

Computational-Level Accounts of RT

The principle example of this approach was to add an assumption called the RT–distance hypothesis to GRT, which simply assumes that RT decreases with the distance between the percept and the decision bound. This assumption was first investigated in SDT (e.g., Murdock, 1985). The idea is that if decisions are made by comparing a percept to a decision bound or criterion, then the greater the distance between the two, the easier, and hence the faster, the decision. This simple assumption has received considerable empirical support (Ashby, Boynton, & Lee, 1994; Murdock, 1985). As noted earlier, Ashby and Maddox (1994) showed that if the RT–distance hypothesis holds, then strong nonparametric RT tests of perceptual separability are possible.

Process Models of RT

This has been the more popular approach. Ashby (2000) generalized the drift-diffusion model described earlier to multiple perceptual dimensions. In this version, the perceptual representations are the same as in classical GRT. Like the drift-diffusion model, application was restricted to tasks with two response alternatives. On each trial, the observer's experience with the stimulus was assumed to produce repeated samples from the relevant perceptual distribution. Each sample is compared to the decision bound and a signed distance is computed, which equals distance-to-bound if the percept is in the A region and minus distance-to-bound if it is in the B region. At this point, the model is identical to the drift-diffusion model – that is, the signed distances are cumulated, and sampling continues until the sum crosses an upper or lower barrier (exactly as in Figure 6.5, except with an "A" response replacing "YES" and a "B" response

replacing "NO"). Ashby (2000) showed that this model includes the static version of GRT as a special case, and showed that the variance–covariance matrices estimated in classical applications of GRT are corrupted by decisional influences. For example, consider two conditions in which the task is identical but participants are pressed to respond more quickly in one than the other. In general, we expect more errors in the condition with speed stress. Fitting the static GRT model to these data would suggest that perceptual noise increases with speed stress. In contrast, the stochastic version of GRT accounts for these data by reducing the distance to the response barriers in the speeded condition (i.e., the numerical values of A and B in Figure 6.5), but not changing perceptual noise.

More recently, P. L. Smith (2019) proposed a similar model, except based on a circular diffusion process. The model can be applied to a variety of different tasks, but consider its application to the 2×2 factorial identification experiment with stimuli A_1B_1, A_1B_2, A_2B_1, and A_2B_2. As mentioned earlier, in static GRT models of this task, the origin of the perceptual space is arbitrary. Suppose we define the origin as the center point of the four perceptual means (i.e., the mean of the means), and the drift is determined by cumulating random samples from the perceptual distribution associated with the presented stimulus (e.g., scaled by some multiplicative constant). Then the drift will generally be outwards and in the direction of the perceptual mean of the stimulus. P. L. Smith (2019) assumed a single circular absorbing barrier that is divided into four quadrants – one associated with each response alternative. The accumulation process continues until absorption occurs, at which point the associated response is given. A response bias toward or against a particular response can be implemented by setting the angle of the response quadrant associated with that response to be greater or less than 90°, respectively. Because this task includes more than two response alternatives, the stochastic GRT model proposed by Ashby (2000) is not even defined in this case. So in this sense, Smith's model has a considerable advantage over the model proposed by Ashby. On the other hand, the circular-diffusion model does not include decision bounds, so it is unclear how the model would account for performance differences that arise, for example, when the participant switches to or away from bounds that satisfy decisional separability.

As noted earlier, Townsend and colleagues (Townsend, Houpt, & Silbert, 2012; Townsend & Wenger, 2004; Wenger & Townsend, 2006) interpreted GRT within the framework of stochastic linear dynamical systems. These models assume the stimulus dimensions or components are processed by parallel channels that are potentially interactive (Townsend *et al.*, 2020). Activation in each channel is accumulated until it reaches a criterion level, and the outputs of the different channels are then passed to decisional operators (e.g., Boolean AND or OR gates). Like the drift-diffusion interpretations of GRT, these models make simultaneous accuracy and RT predictions. They have also been used to model configurality (Wenger & Townsend, 2006) and to derive new RT summary statistics that can be used to test for perceptual separability and perceptual independence.

6.4.6 Extensions to Neuroscience

GRT was developed before the cognitive neuroscience revolution that began in the 1990s. As a result, for its first several decades of existence, GRT was a purely perceptual and cognitive theory. But during the past several decades there has been progress on two fronts. First, much has been learned about the architecture and functioning of the neural circuits that implement the perceptual and decision processes hypothesized by GRT. Second, GRT analyses have recently been extended to neuroscience data, in particular to data from neuroimaging experiments. This section briefly reviews these trends. For more details, see Ashby and Soto (2016) and Soto, Vucovich, and Ashby (2018).

There is now overwhelming evidence that humans have multiple learning systems that for the most part are neuroanatomically and functionally distinct (e.g., Ashby & Maddox, 2005; Eichenbaum & Cohen, 2001; Squire, 2004). The most complete description of two of the most important learning systems is arguably provided by the COVIS theory (Ashby & Valentin, 2017; Ashby *et al.*, 1998). COVIS assumes separate explicit-reasoning and procedural-learning systems that compete for access to response production. The explicit-reasoning system uses executive attention and working memory to learn explicit rules, and is mediated by a broad neural network that includes the prefrontal cortex, anterior cingulate, head of the caudate nucleus, and the hippocampus. In contrast, the procedural system uses dopamine-mediated reinforcement learning when the optimal strategy is difficult or impossible to describe verbally, and key structures include the striatum and premotor cortex.

Knowing which learning system participants are using can facilitate a subsequent GRT analysis because the explicit system is constrained to use bounds that satisfy decisional separability (at least locally), whereas the procedural system is not. The explicit system learns and applies explicit rules that can be described using Boolean algebra. More specifically, it makes independent decisions about the level of the stimulus (e.g., high vs. low) on one or more dimensions and then combines the outcomes of these separate decisions using simple logical operators, such as "and" to produce conjunction rules and "or" to produce disjunctions. When translated into decision bounds, the resulting response regions can always be separated by piecewise linear bounds, in which each piece is a vertical or horizontal line segment. Thus, each piece satisfies decisional separability. In contrast, the procedural system implements less constrained decision strategies that are compatible with any of the decision bounds that are used when fitting GRT models. For these reasons, if decisional separability is assumed, then it is vital to select experimental conditions that favor explicit reasoning over procedural learning.

Soto *et al.* (2018) extended GRT analysis to neuroimaging data in the context of a study examining the relationship between facial identity and perceived emotion. When a visual stimulus is presented to an observer, it causes activation in many areas within the visual system. The perceptual representation modeled in GRT

likely depends on activation in some higher-level visual area. If this representation violates perceptual separability (or perceptual independence), then an obvious and important question is when and where separability (or independence) was first violated within the processing stream? To address this question, Soto *et al.* (2018) first defined the concepts of *encoding separability* and *encoding independence*. If a stimulus dimension is encoded in some brain region of interest in exactly the same way when an irrelevant dimension is varied, then the former shows encoding separability from the latter. Similarly, if the neural representations of two stimulus dimensions are statistically independent in some region of interest, then they satisfy encoding independence. Next, Soto *et al.* (2018) proposed empirical tests of these constructs that are based on summary statistics derived from applying pattern classifiers to fMRI data. For example, the first step might be to construct a support vector machine that classifies the level of stimulus dimension A in some brain region of interest as 1 or 2 (following methods described, e.g., by Ashby, 2019). *Decoding separability* holds if the distributions of decoded values of dimension A are invariant across changes in a second, irrelevant dimension B.[11]

Similarly, Wenger and Rhoten (2020) demonstrated that it was possible to use the timing of a feature in EEG data to draw inferences regarding independence and separability in a study of visual perceptual learning. Specifically, they used the onset time of the lateralized readiness potential (LRP). The LRP is a negative-going waveform, measured in central electrodes contralateral to the motor response that it precedes, and is interpreted as indicating that sufficient processing has been completed in order to program the motor response. The onset time of the LRP was shown to be strongly correlated with observable RT. Consequently, when those onset times were analyzed with respect to timed marginal response invariance and timed report independence (see the subsection entitled "Summary Statistics Approach"), they were found to support inferences that were consistent with the inferences drawn from the response frequencies.

6.5 Concluding Remarks

The power and generality of statistical decision theory – SDT in one dimension and GRT in multiple dimensions – should confirm Estes' evaluation that SDT is ". . . the most towering achievement of basic psychological research in the last half century" (Estes, 2002, p. 15). One would be hard-pressed to name a sub-discipline of the behavioral sciences (cognitive neuroscience included) that does not concern themselves with aspects of identification and categorization (classification). This fact, along with the fact that SDT "scales" to dealing with

11 Operationally, this can be tested in the following way. Consider an identification experiment with stimuli A_1B_1, A_1B_2, A_2B_1, and A_2B_2. First, compute the distance of each activity vector to the classifier hyperplane. Second, estimate the distributions of the A_1 and A_2 distances separately when B is at level 1 and at level 2. Finally, compare the A_1 distributions when B is at level 1 and at level 2, and also compare the A_2 distributions when B is at level 1 and level 2.

neurophysiological data, perhaps reinforces Wixted's opinion that "... it should not be possible to earn a Ph.D. in experimental psychology without having some degree of proficiency in signal detection theory" (Wixted, 2020, p. 225). Along with these kinds of advances, we should note that a critical strength of the community of researchers associated with SDT and GRT is the unflinching willingness to tackle difficult problems, such as the identifiability issues discussed here. Investigators have added and continue to develop novel and improved methods for framing hypotheses and connecting theory and data.

6.6 Related Literature

Link (1994) and Wixted (2020) provide excellent historical overviews of the antecedents to SDT and to its early years. The original classic text on SDT was by Green and Swets (1966). It remains relevant today, especially for its treatment of ideal observer theory. For more recent texts, see Macmillan and Creelman (2005) or Wickens (2002).

There is no text on GRT, although this topic is briefly covered by Macmillan and Creelman (2005). Even so, there are a few recent GRT tutorials, including by Ashby and Soto (2015) and Silbert and Hawkins (2016). For a review of the mathematical foundations of GRT, see Fukunaga (2013).

Acknowledgments

We thank Matthew Crossley and Jeffrey Inglis for their helpful comments on this chapter.

References

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology*, *44*(2), 310–329.

Ashby, F. G. (2019). *Statistical analysis of fMRI data*, 2nd ed. Cambridge, MA: MIT Press.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, *55*(1), 11–27.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*(4), 423–466.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124–150.

Ashby, F. G., & Soto, F. A. (2015). Multidimensional signal detection theory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 13–34). New York: Oxford University Press.

Ashby, F. G., & Soto, F. A. (2016). The neural basis of general recognition theory. In *Mathematical models of perception and cognition: Volume II, A Festschrift for James T. Townsend* (pp. 1–31). New York: Routledge.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science*, 2nd ed. (pp. 157–188). New York: Elsevier.

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience*, 4th ed. *(Volume 5: Methodology)* (pp. 1–41). New York: Wiley.

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, *130*(1), 77–96.

Balakrishnan, J. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(5), 1189–1206.

Bridgman, P. W. (1945). Some general principles of operational analysis. *Psychological Review*, *52*(5), 246–249.

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an *n*-way generalization of Eckart–Young decomposition. *Psychometrika*, *35*(3), 283–319.

Cornes, K., Donnelly, N., Godwin, H., & Wenger, M. J. (2011). Perceptual and decisional factors affecting the detection of the Thatcher illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 645–668.

Creelman, C. D. (2015). Signal detection theory, history of. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences*, 2nd ed. (pp. 952–956). New York: Elsevier.

Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1388–1403.

DeValois, R. L., & De Valois, K. K. (1990). *Spatial vision*. New York: Oxford University Press.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *38*(4), 840–859.

Dunnington, G. W., Gray, J., & Dohse, F. E. (2004). *Carl Friedrich Gauss: Titan of science*. Washington, DC: Mathematical Association of America.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain*. New York: Oxford University Press.

Ennis, D. M., & Ashby, F. G. (2003). Fitting decision bound models to identification or categorization data. Unpublished manuscript. Available at https://www.researchgate.net/profile/F-Ashby/publication/255572437_Fitting_Decision_Bound_Models_to_Identification_or_Categorization_Data/links/54f9e1f80cf29a9fbd7c5740/Fitting-Decision-Bound-Models-to-Identification-or-Categorization-Data.pdf..

Estes, W. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*(1), 3–25.

Fechner, G. T. (1860). *Elements of psychophysics* (translated by H. E. Adler, 1966). Leipzig: Breitkopf & Härtel (Holt, Rinehart, & Winston).

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(1), 69–78.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*, 2nd ed. San Diego, CA: Academic Press.

Fullerton, G., & Cattell, J. (1892). *On the perception of small differences*. University of Pennsylvania Philosophy Series, No. 2. Philadelphia, PA: University of Pennsylvania.

Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*(3), 225–241.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*(3), 149–159.

Garner, W. R., & Morton, J. (1969). Perceptual independence: Definitions, models, and experimental paradigms. *Psychological Bulletin*, *72*, 233–259.

Green, D. M. (1964). General prediction relating yes–no and forced-choice results. *The Journal of the Acoustical Society of America*, *36*(5), 1042.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Helmholtz, H. v. (1867). *Handbuch der physiologischen Optik* (vol. 9). Leipzig: Voss.

Kadlec, H. (1995). Multidimensional signal detection analyses (MSDA) for testing separability and independence: A PASCAL program. *Behavior Research Methods, Instruments, & Computers*, *4*, 442–458.

Kadlec, H., & Townsend, J. T. (1992a). Implications of marginal and conditional detection parameters for the separabilities and independence of perceptual dimensions. *Journal of Mathematical Psychology*, *36*, 325–374.

Kadlec, H., & Townsend, J. T. (1992b). Signal detection analysis of dimensional interactions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 181–228). Hillsdale, NJ: Erlbaum.

Killeen, P. R., & Taylor, T. J. (2004). Symmetric receiver operating characteristics. *Journal of Mathematical Psychology*, *48*(6), 432–434.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher

and Neyman–Pearson. *The British Journal for the Philosophy of Science*, *57*(1), 69–91.

Link, S. W. (1994). Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science*, *5*(6), 335–340.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*(1), 77–105.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology,* (Vol. 1, pp. 103–190). New York: Wiley.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*(5), 945–952.

Maddox, W. T., & David, A. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100–107.

Marcum, J. I. (1947). *A statistical theory of target detection by pulsed radar* (Tech. Rep.). Santa Monica, CA: Rand Corporation.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.

Menneer, T., Wenger, M., & Blaha, L. (2010). Inferential challenges for general recognition theory: Mean-shift integrality and perceptual configurality. *Journal of Vision*, *10*(7), 1211–1211.

Murdock, B. B. (1985). An analysis of the strength-latency relationship. *Memory & Cognition*, *13*(6), 511–521.

Murphy, J., Gray, K. L. H., & Cook, R. (2017). The composite face illusion. *Psychonomic Bulletin & Review*, *24*(2), 245–261.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694–706), 289–337.

O'Toole, A. J., Wenger, M. J., & Townsend, J. T. (2001). Quantitative models of perceiving and remembering faces: Precedents and possibilities. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 1–38). Mahwah, NJ: Erlbaum.

Peterson, W. W., & Birdsall, T. G. (1953). *The theory of signal detectability: Part I. The general theory.* Electronic Defense Group Technical Report 13, University of Michigan.

Peterson, W. W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 171–212.

Piepers, D., & Robbins, R. (2012). A review and clarification of the terms "holistic," "configural," and "relational" in the face perception literature. *Frontiers in Psychology*, *3*, 559.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254–276.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*, *140*(5), 1281–1302.

Richler, J. J., Gauthier, I., Wenger, M. J., & Palmeri, T. J. (2008). Holistic processing of faces: Perceptual and decisional components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 328–342.

Rose, A. (1942). The relative sensitivities of television pickup tubes, photographic film, and the human eye. *Proceedings of the IRE*, *30*(6), 293–300.

Rose, A. (1948). The sensitivity performance of the human eye on an absolute scale. *Journal of the Optical Society of America*, *38*(2), 196–208.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139–253.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Silbert, N. H., & Hawkins, R. X. D. (2016). A tutorial on general recognition theory. *Journal of Mathematical Psychology*, *73*, 94–109.

Silbert, N. H., & Thomas, R. D. (2013). Decisional separability, model identification, and statistical inference in the general recognition theory framework. *Psychonomic Bulletin & Review*, *20*(1), 1–20.

Silbert, N. H., & Thomas, R. D. (2017). Identifiability and testability in GRT with individual differences. *Journal of Mathematical Psychology*, *77*, 187–196.

Smith, J. K. (1992). Alternative biased choice models. *Mathematical Social Sciences*, *23*(2), 199–219.

Smith, P. L. (2019). Linking the diffusion model and general recognition theory: Circular diffusion with bivariate-normally distributed drift rates. *Journal of Mathematical Psychology*, *91*, 145–158.

Soto, F. A., Vucovich, L., Musgrave, R., & Ashby, F. G. (2015). General recognition theory with individual differences: A new method for examining perceptual and decisional interactions with an application to face perception. *Psychonomic Bulletin & Review*, *22*(1), 88–111.

Soto, F. A., Vucovich, L. E., & Ashby, F. G. (2018). Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data. *PLoS Computational Biology*, *14*(10), e1006470.

Soto, F. A., Zheng, E., Fonseca, J., & Ashby, F. G. (2017). Testing separability and independence of perceptual dimensions with general recognition theory: A tutorial and new R package (grtools). *Frontiers in Psychology*, *8*, 696.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Tanner, W. (1956). Theory of recognition. *The Journal of the Acoustical Society of America*, *28*(5), 882–888.

Thomas, R. D. (2001). Characterizing perceptual interactions in face identification using multidimensional signal detection theory. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 193–228). Mahwah, NJ: Erlbaum.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Thurstone, L. L. (1927b). Psychophysical analysis. *The American Journal of Psychology*, *38*(3), 368–389.

Townsend, J. T., Houpt, J. W., & Silbert, N. H. (2012). General recognition theory extended to include response times: Predictions for a class of parallel systems. *Journal of Mathematical Psychology*, *56*(6), 476–494.

Townsend, J. T., Liu, Y., Zhang, R., & Wenger, M. J. (2020). Interactive parallel models: No Virginia, violation of Miller's race inequality does not imply coactivation and yes Virginia, context invariance is testable. *The Quantitative Methods for Psychology*, *16*(2), 192–212.

Townsend, J. T., & Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, 1003–1035.

Townsend, J. T., & Wenger, M. J. (2015). On the dynamic perceptual characteristics of Gestalten: Theory-based methods. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 948–968). Oxford: Oxford University Press.

Treutwein, B., Rentschler, I., & Caelli, T. (1989). Perceptual spatial frequency-orientation surface: Psychophysics and line element theory. *Biological Cybernetics*, *60*(4), 285–295.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026.

Van Meter, D., & Middleton, D. (1954). Modern statistical approaches to reception in communication theory. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 119–145.

Von Der Heide, R. J., Wenger, M. J., Bittner, J. L., & Fitousi, D. (2018). Converging operations and the role of perceptual and decisional influences on the perception of faces: Neural and behavioral evidence. *Brain and Cognition*, *122*, 59–75.

Wenger, M. J., & Ingvalson, E. M. (2002). A decisional component of holistic encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 872–892.

Wenger, M. J., & Ingvalson, E. M. (2003). Preserving informational separability and violating decisional separability in facial perception and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1106–1118.

Wenger, M. J., & Rhoten, S. E. (2020). Perceptual learning produces perceptual objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 455–475.

Wenger, M. J., & Townsend, J. T. (2006). On the costs and benefits of faces and words. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 755–779.

Wickens, T. D. (1992). Maximum-likelihood estimation of a multivariate Gaussian rating model with excluded data. *Journal of Mathematical Psychology*, *36*(2), 213–234.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, 14.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*, 747–759.

7 Modeling Response Inhibition in the Stop-Signal Task

Hans Colonius and Adele Diederich

7.1 Response Inhibition and the Stop-Signal Task

The notion of *response inhibition* refers to an organism's ability to suppress unwanted impulses, or actions and responses that are no longer required or have become inappropriate. This ability is considered a case of cognitive control, those cognitive faculties that allow information processing and behavior to vary adaptively from moment to moment depending on current goals, rather than remaining rigid and inflexible. At this time, the field of cognitive control flourishes like never before (Logan, 2017, p. 875).[1]

The simple fact that cognitive control takes time makes subjects' behavior amenable to the advanced methods of response time analysis and modeling developed in cognitive psychology over many years. In the *stop-signal paradigm*, participants typically perform a *go task* (e.g., press left when an arrow pointing to the left appears, and press right when an arrow pointing to the right appears), but on a minority of the trials, a stop signal (e.g., an acoustic stimulus) appears after a variable stop-signal delay, instructing the participant to suppress the imminent go response (see Figure 7.1). This paradigm has become the main workhorse being used in laboratory settings across various human populations (e.g., clinical vs. nonclinical, different age groups) as well as nonhuman ones (primates, rodents, etc.).

The stop-signal task provides three types of observable data: (i) reaction times (RTs) to the go signal in go trials; (ii) RTs in stop trials (when response inhibition failed); and (iii) the frequency of responses given in spite of the stop signal. Unlike the latency of go responses, response-inhibition latency cannot be observed directly (as successful response inhibition results in the absence of an observable response). This is a problem, in particular because the time to cancel a response is widely considered to be an appropriate indicator of the level of response inhibition of an individual, and it must be addressed by any model of the stop-signal task.

The main goal of this chapter is to present results in the formal modeling of behavioral data from the stop-signal paradigm and some of its variants. Given that there exist a number of comprehensive literature reviews of both empirical and modeling results (see Section 7.10 on related literature), we primarily present a general formal framework allowing us to incorporate most current models and, at the same time, expose a number of open or only partially solved problems. In order to keep the chapter self-contained, we start by presenting some typical

1 Gordon Logan emphasizes that "Cognitive control addresses core issues in basic and applied psychology, from free will and the nature of intention to practical strategies for improving our own control and treating deficient control in our clients." According to Web of Science (10/2020), there were about 200 papers and 10,000 citations in 2019 for "stop-signal task."

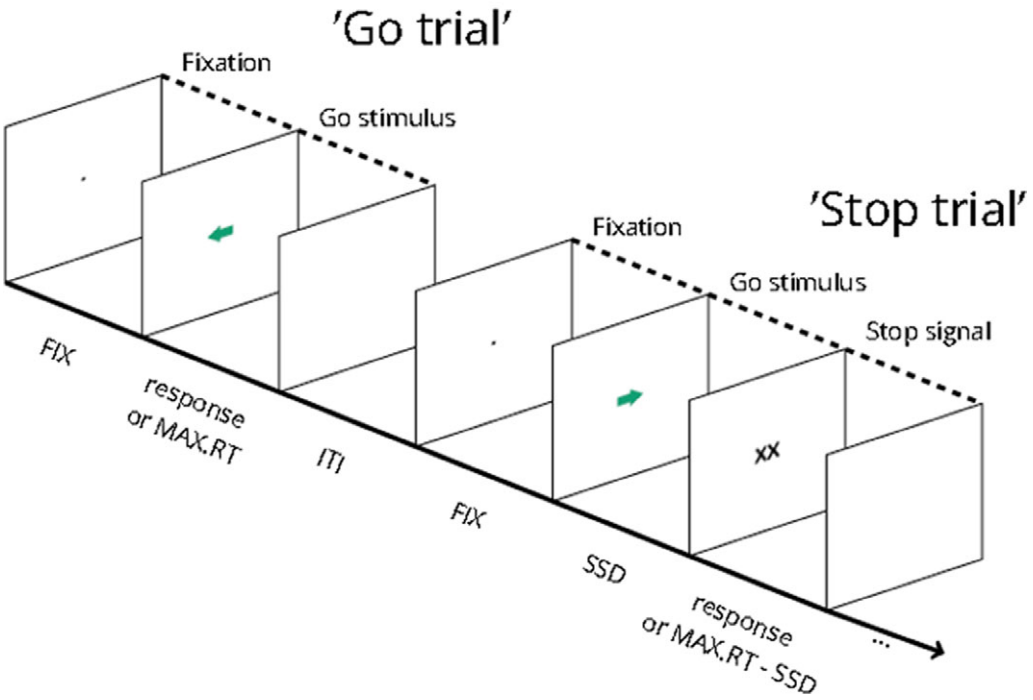


Figure 7.1 *Depiction of the sequence of events in a stop-signal task. In this example, participants respond to the direction of the small arrows (by pressing the corresponding arrow key) in the go task. On a quarter of the trials, the arrow is replaced by "XX" after a variable stop-signal delay (FIX = fixation duration; SSD = stop-signal delay; MAX.RT = maximum reaction time; ITI = intertrial interval) (from Verbruggen* et al.*, 2019).*

data patterns. Then, the general race model is introduced including estimation methods for the nonparametric case (Section 7.3). More detailed presentations of parametric independent (Section 7.4) and dependent (Section 7.5) race models follow. Some related, but non-race, models are discussed in Section 7.6. Section 7.7 introduces the class of semi-parametric race models based on the copula concept. It also contains the race model with perfect negative dependence. Variants of the stop-signal paradigm, the problems of trigger failures, and sequential effects are sketched in Section 7.8. We conclude with a brief discussion contrasting parametric versus nonparametric approaches and a look into the future of stop-signal modeling. A list of abbreviations used in the chapter is found in Table 7.1.

Table 7.1 *Abbreviations used in the chapter.*

Acronym	Meaning
ARI	anticipated response inhibition
CI	context independence
FEF	frontal eye field
FGM	Farlie–Gumbel–Morgenstern (copula)
IT	inhibition time
LATER	linear approach to threshold with ergodic rate (model)
MCMC	Markov chain Monte Carlo
PND	perfect negative dependency
PTC	pause-then-cancel (model)
RT	response (or reaction) time
SC	superior colliculus
SI	saccadic inhibition
SOA	stimulus-onset asynchrony
SSD	stop-signal delay
SSRT	stop-signal reaction time

7.2 Some Typical Data Patterns in the Stop-Signal Paradigm

Given the popularity of the stop-signal task, the amount of data is enormous and, unsurprisingly, there is a lot of diversity in the findings due to differences in design, instruction, and the specific subpopulation tested. Nonetheless, many results only differ with respect to their specific numerical values observed for reaction times and inhibition probabilities, while some general qualitative features of the inhibition function and RT distributions are typically retained.

7.2.1 Inhibitions Function

Inhibitions functions depict the probability of a response in spite of a stop signal as a function of stop-signal delay (SSD).[2] When the stop signal occurs soon after the go signal, participants have a high chance of withholding a response, so the inhibition function has a small value. With SSD increasing, this chance diminishes more and more, up to a point where the probability to respond approaches 1. The top panel of Figure 7.2 depicts classic data from three subjects reported in Logan and Cowan (1984). While these inhibition functions are somewhat similar in shape, subjects clearly differ: for mid-range SSD values, the probability of a response can vary enormously. Does this imply that, e.g., participant J.M. (lower curve) is much better in controlling the response than the other two? Unfortunately, interpretation of inhibition functions is not straightforward. Although J.M.'s

2 Strictly speaking, it should be called "non-inhibition function," but the terminology used here is common.

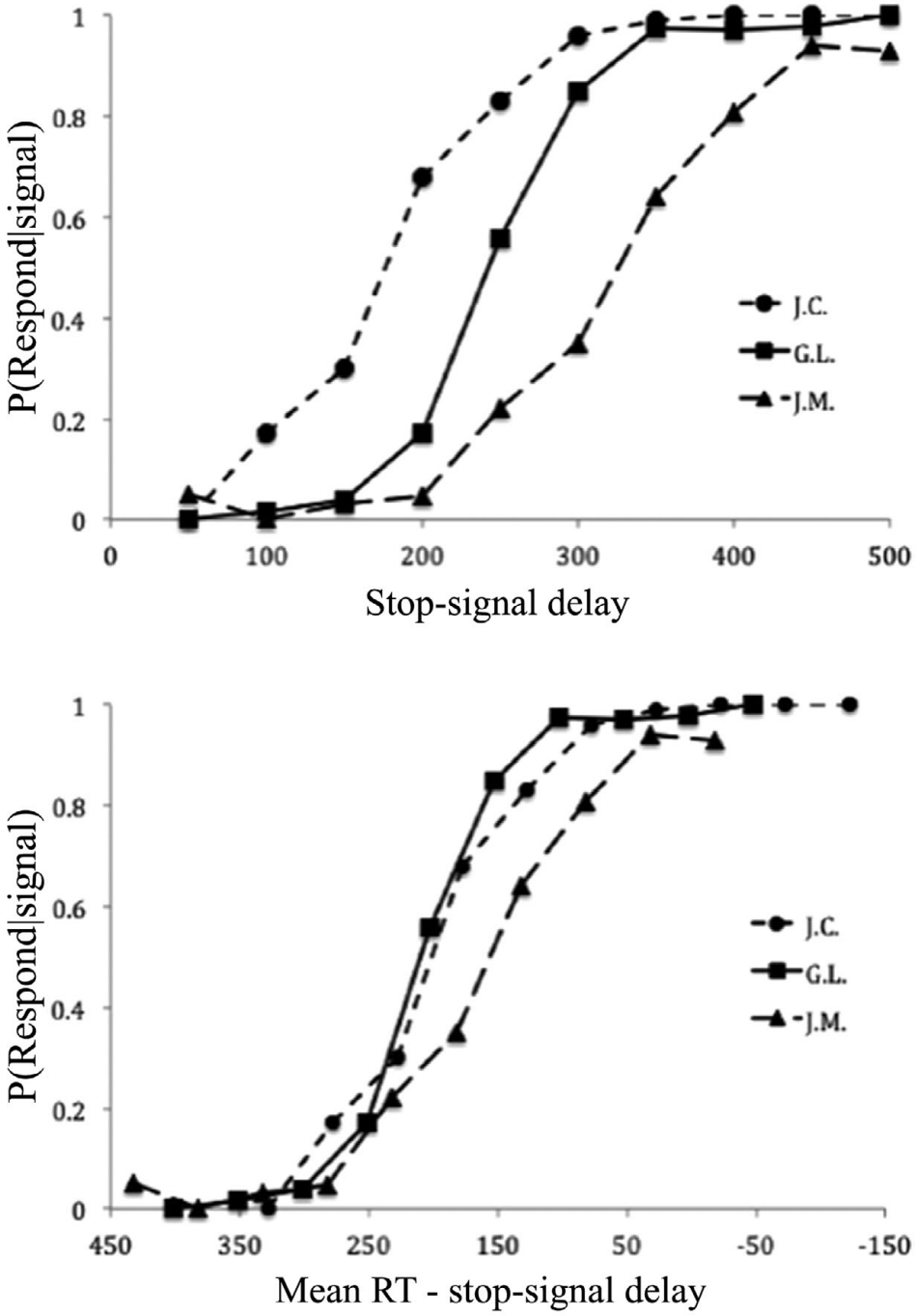


Figure 7.2 *Classic data reported in Logan and Cowan (1984). Top panel: Inhibition functions from three subjects plotted as a function of stop-signal delay. Bottom panel: Inhibition probability for the same three subjects replotted as a function of mean go response time minus stop-signal delay (SSD) (from Logan* et al.*, 2014).*

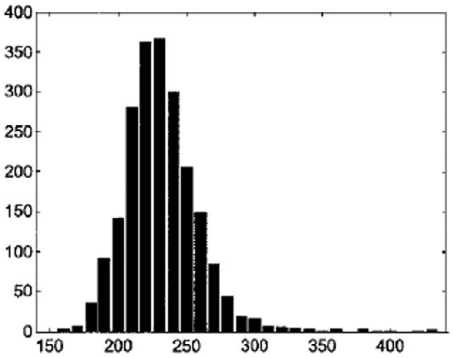


Figure 7.3 *Saccadic reaction times to a visual target with* $N = 2144$ *(from Özyurt, Colonius, & Arndt, 2003, subject P.T.).*

inhibitory performance may in fact be best, it could also be the result of J.M. voluntarily slowing responses to the go signal in most trials, so that there is always "enough" time to stop the response. Even if one persuades participants to not delay their response, it has still been shown that various parameters of the distribution of responses to the go signal, like variance, may have a strong effect on the inhibition functions. Suggestions to remove these problems by a standardized transformation of the inhibition function remain controversial, however. The bottom panel of Figure 7.2 shows the probability of inhibition plotted against the difference between mean go RT and stop-signal delay for the same subjects. Under some simplifying assumptions, this difference is interpretable as a measure of the time that is available to detect the stop signal and to cancel a response.[3] In sum, this issue calls for developing a formal model within which the level of performance can be gauged exactly by some parameter estimated from the data.

7.2.2 Reaction Times to Go and Stop Signal

The distribution of reaction times on go trials (i.e., without a stop signal) is often more or less right-skewed, as is typical for RT distributions in general. Figure 7.3 depicts the histogram of 2144 saccadic reaction times to a visual target, occurring either to the left or right of the fixation point, in a stop-signal task with auditory stop signals (Özyurt, Colonius, & Arndt, 2003). Responses on unsuccessful stop trials (*signal-respond* RT) are on average faster than go RT on trials with no stop signal and faster for shorter stop-signal delays than for longer ones. Note that this latter observation is to be expected assuming that the process of inhibition evolves over a possibly variable time interval. This feature (often called "fan effect"), illustrated in Figure 7.4 by another study on saccadic RTs to a visual target with an auditory stop signal (Colonius, Özyurt, & Arndt, 2001), motivated the development of so-called race models to be discussed below.

3 Note that the functions for J.C. and G.L. align better than the function for J.M. because J.M. had greater variability in go RT than the other two.

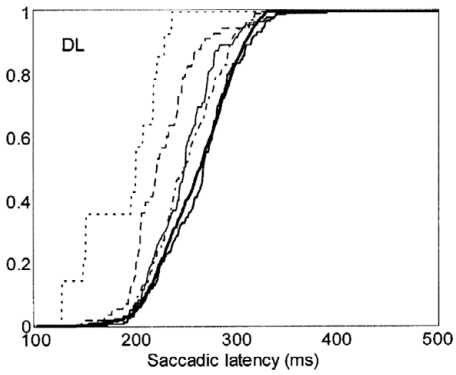


Figure 7.4 *Empirical (signal-respond) distribution functions of saccadic RTs to a visual target with an auditory stop signal presented at different SSD values [ms] (in parentheses: number of observations). Dotted line: 150 (14); dashes: 200 (58); thin line: 230 (122); dots/dashes: 250 (147); medium line: 270 (170); thick line: go condition (2919) (from Colonius, Özyurt, & Arndt, 2001, subject D.L.).*

7.3 Modeling the Stop-Signal Task

Many features of typical data in the stop-signal task are consistent with modeling responses as the outcome of a race between processing of the go signal and the stop signal: if the latter terminates earlier than the former, subjects succeed in inhibiting a response, otherwise they respond in spite of the stop signal. Although "race" is the predominant modeling approach, let us first take a step back and consider the situation from a more general point of view.

When only the go signal is presented, denoted as *context GO*, reaction time T_{go}, say, represents the time to process that signal, including possible pre- and motor components. In order to account for some variability across trials, T_{go} is considered a random variable taking on non-negative values. In contrast, when the go signal is followed by presentation of a stop signal, denoted as *context STOP*, the two alternative outcomes – either a response is given or there is no response – are the result of (somehow) processing both signals. Race models hold that, in addition to T_{go}, there is a separate random processing time for the stop signal, T_{stop}, say, and the outcome is determined by $\min\{T_{go}, T_{stop} + \text{SSD}\}$. Alternatively, instead of claiming a separate processing time T_{stop}, one could assume that the stop signal modulates processing of the go signal in a way that is qualitatively consistent with two fundamental empirical observations. First, RTs in context *STOP* tend to be faster than in context *GO*; thus, according to this alternative view, the stop signal speeds up processing time T_{go} for the go signal. Second, the probability of inhibition decreases with SSD; thus, the later the stop signal is presented, the shorter the time it can modify processing time T_{go}. We will sketch such non-race models in Section 7.6.

Nonetheless, a glance over the stop-signal literature strongly suggests that the race model is the "main game in town," especially when certain generalizations

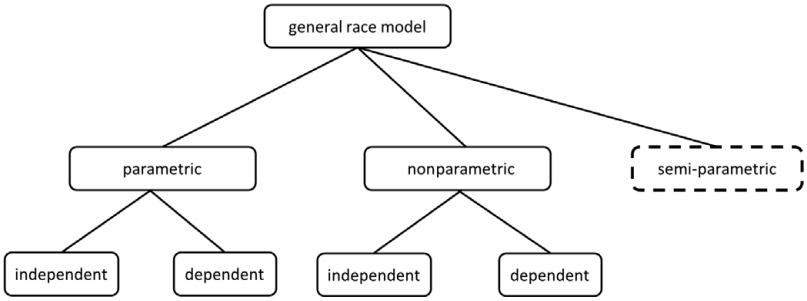


Figure 7.5 *Parametric, nonparametric, and semi-parametric subclasses of the general race model with either independent or dependent T_{go} and T_{stop} processing times.*

and extensions of the notion of "race" are included, like interdependent processing or certain across-trial strategies for optimizing responses. Therefore, the chapter's focus is on this model class. The next section provides a formal introduction of the race model and its subclasses.

7.3.1 The General Race Model

One important distinction in classifying race models is whether they are parametric or nonparametric, that is, if specific distributional assumptions concerning T_{go} and T_{stop} are made. Another is whether these random variables are considered to be statistically independent or not (see Figure 7.5). Although semi-parametric race models actually contain both parametric and nonparametric instances, they are listed here as a separate subclass for conceptual reasons. They are based on the definition of a copula and will be discussed in Section 7.7.

For context *STOP*, we postulate a bivariate cumulative distribution function (cdf), denoted H, for T_{go} and T_{stop}:

$$H(s,t) = \mathrm{P}\left[T_{go} \leq s, T_{stop} \leq t\right], \tag{7.1}$$

defined for all real numbers s and t, with $s,t \geq 0$. Moreover, T_{go} and T_{stop} are assumed to be continuous random variables.[4] Sometimes T_{stop} is referred to as *stop-signal reaction time* (SSRT). The marginal cdfs of $H(s,t)$ are denoted as

$$F_{go}(s) = \mathrm{P}\left[T_{go} \leq s, T_{stop} < \infty\right] \text{ and}$$
$$F_{stop}(t) = \mathrm{P}\left[T_{go} < \infty, T_{stop} \leq t\right].$$

In context *STOP*, the go signal triggers realization of random variable T_{go} and the stop signal triggers realization of random variable T_{stop}. In context *GO*, however, only processing of the go signal occurs. Thus, the two different experimental conditions in the paradigm, *GO* and *STOP*, imply the existence of two different

4 That is, H possesses a bivariate density.

sample spaces in the statistical modeling of the task. In principle, the distribution in context *GO*, $F^*_{go}(s)$, say, could be different from the marginal distribution $F_{go}(s)$ in context *STOP*.

However, the general race model rules this out by adding the important assumption of *context independence*, also known as *context invariance*.[5]

Context independence (CI) In context *GO*, the distribution of go-signal processing time is assumed to be

$$F^*_{go}(s) \equiv F_{go}(s) = \mathrm{P}\,[T_{go} \leq s, T_{stop} < \infty] \tag{7.2}$$

for all s, i.e., it is identical to the marginal distribution $F_{go}(s)$ in context *STOP*.

Note that, in order to be more precise, context *STOP* would have to be indexed by the specific value of SSD, t_d, say, with $t_d \geq 0$, and the same holds for $H(s,t)$ and $F_{stop}(t)$. In the following, however, we will tacitly assume that SSD *invariance* holds, meaning that we can drop the index t_d throughout without consequences while keeping it as a given (design) parameter. Moreover, T_{stop} is set equal to zero for $t \leq t_d$, with probability one.

From these assumptions, the probability of observing a response (r) to the go signal given a stop signal was presented with SSD $= t_d$ [ms] after the go signal, is defined by the *race assumption*

$$p_r(t_d) = \mathrm{P}\,[T_{go} < T_{stop} + t_d]. \tag{7.3}$$

In addition, according to the model, the probability of observing a response to the go signal no later than time t, given the stop signal was presented with delay t_d, is given by the (conditional) distribution function

$$F_{sr}(t \mid t_d) = \mathrm{P}\,[T_{go} \leq t \mid T_{go} < T_{stop} + t_d], \tag{7.4}$$

also known as *signal-respond RT (sr)* distribution.

The main interest in modeling the race is to derive information about the distribution of the non-observable stop signal processing time, T_{stop}, or about some of its parameters given sample estimates of $F_{go}(t)$, $F_{sr}(t \mid t_d)$, and $p_r(t_d)$. For example, the independent race model presented in Section 7.3.2 is parameter-free, i.e., no parameters have to be estimated in order to make predictions. Later, we will discuss both fully parameterized models and semi-parametric versions. In the latter, no specific distributions are postulated but only a parameter assessing the degree of stochastic dependency.

The most simple version of the race model, sometimes referred to as *independent horse race model*, assumes the non-observable time $T_{stop} =$ SSRT to be a constant k, $k \geq 0$. Thus, $p_r(t_d)$ becomes simply

$$p_r(t_d) = \mathrm{P}\,[T_{go} \leq t_d + k].$$

5 *Context invariance* seems a more fitting term but to avoid confusion, we keep the familiar *context independence*.

Figure 7.5 is the standard depiction of this specific model. It illustrates how the probability to respond given a stop signal (the area under the curve to the left of the vertical line) depends on (i) SSD (thus generating the inhibition function), (ii) the go RT distribution, and (iii) the stop-signal processing time (SSRT).

Assuming constant stop-signal processing time is not realistic and may impair model predictions (Verbruggen & Logan, 2009), but it simplifies estimation of SSRT enormously. In fact, a popular estimation method for SSRT, the *integration method*, requires it (see below).

7.3.2 The (Complete) Independent Race Model

The most common version of the race model is the (complete) independent race model[6] introduced by Logan and Cowan (1984); it postulates stochastic independence between T_{go} and T_{stop}:

Stochastic independence.

$$H(s,t) = \mathrm{P}\left[T_{go} \le s\right] \times \mathrm{P}\left[T_{stop} \le t\right] = F_{go}(s) \times F_{stop}(t), \tag{7.5}$$

for all s,t $(s,t \ge 0)$.

From this, we have

$$\begin{aligned} p_r(t_d) &= \mathrm{P}\left[T_{go} < T_{stop} + t_d\right] \\ &= \int_0^\infty f_{go}(t)[1 - F_{stop}(t - t_d)]\,dt, \end{aligned} \tag{7.6}$$

with $f_{go}(t)$ denoting the probability density function (pdf) for T_{go}. Moreover, the signal-respond distribution is

$$\begin{aligned} F_{sr}(t \mid t_d) &= \mathrm{P}\left[T_{go} \le t \mid T_{go} < T_{stop} + t_d\right] \\ &= \frac{1}{p_r(t_d)} \int_0^t f_{go}(t')[1 - F_{stop}(t' - t_d)]\,dt', \end{aligned} \tag{7.7}$$

for all $t > t_d$ and $p_r(t_d) > 0$.[7]

The predominance of the independent model is due to the fact that its predictions are mostly consistent with the empirical observations presented above. First, increasing t_d in Equation (7.6) monotonically increases the expression under the integral, thus increasing the probability of a response and approaching 1 in the limit

6 The attribute "complete" is sometimes used to distinguish this model from the one with constant SSRT.

7 One can define $F_{sr}(t \mid t_d)$ for $t \le t_d$ as well: Equation (7.7) then results in

$$F_{sr}(t \mid t_d) = \min\left\{\frac{F_{go}(t)}{p_r(t_d)}, 1\right\}.$$

It is the probability of an anticipatory response (given even before the stop signal is presented), but these responses are usually removed.

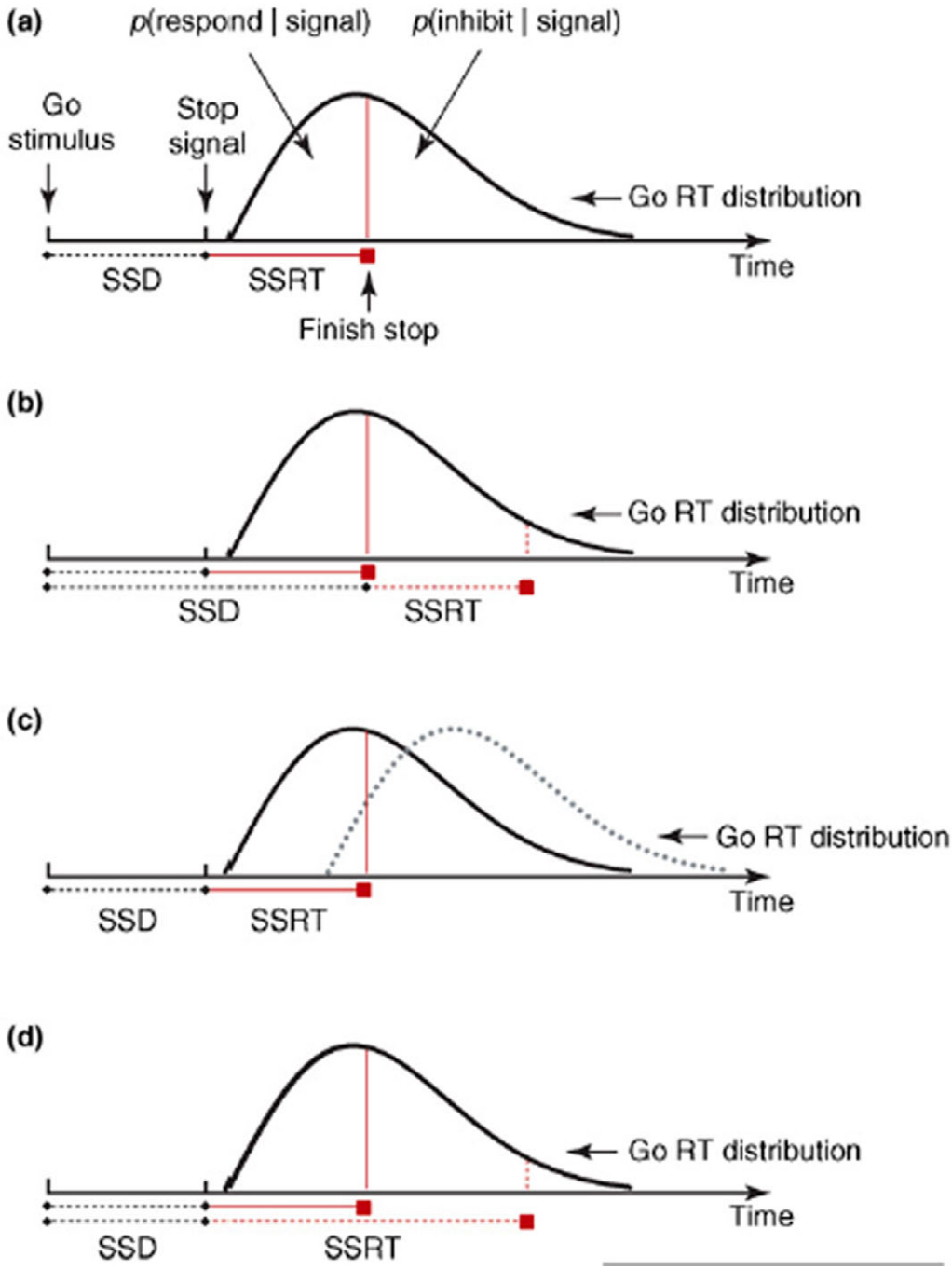


Figure 7.6 *Schematic of the simplified race model: the probability to respond given a stop signal (the area under the curve to the left of the vertical line) depends on t_d = SSD (panel b), go RT distribution (panel c), and stop-signal processing time (SSRT = k) (panel d) (from Verbruggen & Logan, 2008).*

for $t_d \to +\infty$, as observed in Figure 7.2 (top panel). Second, letting $t_d \to +\infty$ in Equation (7.7) implies $F_{sr}(t \mid t_d)$ to approach $F_{go}(t)$, for any fixed t (Figure 7.4). As an additional test, the signal-respond distribution has been shown to have an upper and a lower bound (Colonius, Özyurt, & Arndt, 2001):

$$F_{go}(t) \leq F_{sr}(t \mid t_d) \leq F_{go}(t)/p_r(t_d) \tag{7.8}$$

for all t. The lower bound implies, in particular, that

$$\mathrm{E}[T_{go} \mid T_{go} < T_{stop} + t_d] \leq \mathrm{E}[T_{go}],$$

i.e., mean stop-failure responses should be faster than mean go-signal responses.

Writing $f_{sr}(t \mid t_d)$ for the pdf of $F_{sr}(t|t_d)$, it follows (Colonius, 1990) that

$$f_{sr}(t \mid t_d) = f_{go}(t)\,[1 - F_{stop}(t - t_d)]/p_r(t_d). \tag{7.9}$$

From that, an explicit expression for the distribution of unobservable stop-signal processing time (T_{stop}) follows after rearrangement:

$$F_{stop}(t - t_d) = 1 - \frac{f_{sr}(t \mid t_d)p_r(t_d)}{f_{go}(t)}. \tag{7.10}$$

Unfortunately, simulation studies revealed that gaining reliable estimates for the stop-signal distribution using Equation (7.10) requires unrealistically large numbers of observations (Band, van der Molen, & Logan, 2003; Matzke *et al.*, 2013). As long as one is satisfied with obtaining just an estimate of some parameter of the stop-signal distribution, like the mean, two common "nonparametric" methods are available. If the entire distribution is of interest, a parametric model assuming a distributional family, like the ex-Gaussian, is called for. Both alternatives will be discussed.

7.3.3 Nonparametric Estimation of Stop-Signal Distribution under Independence

We first describe the underlying theoretical assumptions of the methods, followed by some practical considerations for their usage.

Mean method. Rewriting the probability of a response given the stop signal is presented at SSD $= t_d$ as

$$p_r(t_d) = \mathrm{P}\,[T_{go} - T_{stop} < t_d],$$

it can be interpreted formally as the cdf of a random variable T_d, say, taking values t_d, see Logan and Cowan (1984) and illustrated by the shape of Figure 7.2 (bottom panel). It follows that $T_{go} - T_{stop}$ and T_d are *equal-in-distribution*.[8] In particular, we get

8 This means they have the same distribution but are not (necessarily) defined on the same sample space; see Chapter 1 in Volume 1 (p. 10) for definitions.

$$\mathrm{E}[T_{stop}] = \mathrm{E}[T_{go}] - \mathrm{E}[T_d] \tag{7.11}$$

for the mean and

$$\mathrm{Var}[T_{stop}] = \mathrm{Var}[T_d] - \mathrm{Var}[T_{go}] \tag{7.12}$$

for the variance of T_d, the latter following due to stochastic independence of T_{go} and T_{stop}.

Integration method. In contrast to the mean method, here stop-signal processing time is taken to be a constant, t_{stop}, say. Thus

$$F_{stop}(t) = \begin{cases} 0, & \text{if } t < t_d + t_{stop} \\ 1, & \text{if } t \geq t_d + t_{stop} \end{cases}. \tag{7.13}$$

Inserting in Equation (7.6) yields

$$\begin{aligned} p_r(t_d) &= \int_0^\infty f_{go}(t)[1 - F_{stop}(t - t_d)]\, dt \\ &= \int_0^{t_d + t_{stop}} f_{go}(t) dt \\ &= F_{go}(t_d + t_{stop}) \end{aligned} \tag{7.14}$$

and inserting in Equation (7.7) yields

$$\begin{aligned} F_{sr}(t \mid t_d) &= \frac{1}{p_r(t_d)} \int_0^t f_{go}(t')[1 - F_{stop}(t' - t_d)]\, dt' \\ &= \begin{cases} \frac{F_{go}(t)}{F_{go}(t_d + t_{stop})}, & \text{if } t < t_d + t_{stop} \\ 1, & \text{if } t \geq t_d + t_{stop} \end{cases}. \end{aligned}$$

The value of t_{stop} is obtained via Equation (7.14) by determining the quantile of the go-signal distribution, $F_{go}^{-1}(t_d + t_{stop})$, and subtracting the corresponding SSD value t_d (see Figure 7.5).

Some practical considerations. Whether the mean or integration method should be used depends in part on the way the stop-signal delays are set. First, one can simply choose a fixed number of SSDs such that the range of the probability of responding, $p_r(t_d)$, is sufficiently covered. The second method adjusts SSDs dynamically using a tracking procedure (mostly, one-up/one-down), as described, e.g., in Matzke, Verbruggen, and Logan (2018).[9] At convergence, this results in an approximate value of SSD (t_d) such that $p_r(t_d) = 0.5$, but step size should be optimized to avoid slow or no convergence. The tracking method typically results

9 "At the beginning of the experiment, stop-signal delay is set to a specific value (e.g., 250 ms) and is then constantly adjusted after stop-signal trials, depending on the outcome of the race. When inhibition is successful, stop-signal delay increases (e.g., by 50 ms); when inhibition is unsuccessful, stop-signal delay decreases (e.g., by 50 ms)."

in a sufficiently varied set of SSD values so that E[T_{stop}] can be estimated easily by subtracting mean SSD from mean RT on go trials corresponding to Equation (7.11), making the mean method the most popular estimation method.

Applying the integration method with a fixed number of SSDs involves rank-ordering the go RTs for each t_d and selecting the nth go RT where n is the number of go RTs multiplied $p_r(t_d)$. Stop-signal delay is then subtracted to arrive at an estimate of t_{stop} [cf. Equation (7.14)]. Estimates from different stop-signal delays are averaged to arrive at a single estimate for each participant, also when the tracking procedure is being used. Simulation results reported in Verbruggen *et al.* (2019) suggest that the integration method produces the most reliable and least biased nonparametric SSRT estimates under the condition that go omissions (i.e., go trials on which the participant did not respond before the response deadline) and premature responses on unsuccessful stop trials (i.e., responses executed before the stop-signal is presented) should be included in the estimation procedure. Due to numerous recommendations in the literature on how to conduct stop-signal experiments (Logan, 1994; Matzke, Verbruggen, & Logan, 2018; Verbruggen *et al.*, 2019), applying nonparametric race models has become a more or less routine task.

7.4 Parametric Independent Race Models

One reason for adopting a parametric distributional family for go and stop signal processing times is the desire to obtain additional measures of inhibition performance, like variance or skew, in order to differentiate, for example, between clinical subpopulations. Another motive is trying to reveal the mechanisms that implement going and stopping and to predict effects of experimental manipulations on stop-signal performance in the context of a substantive process model of response inhibition. A selected set of independent parametric models will be considered here.

In principle, assuming some parametric form for the distributions of T_{go} and T_{stop} and inserting them into the equations for the go and stop-signal distributions [Equations (7.6)–(7.9)] is straightforward, but obtaining closed-form expressions is often not achievable. The signal-respond distribution can be written as

$$\begin{aligned} F_{sr}(t \mid t_d; \theta_{go}; \theta_{stop}) &= \mathrm{P}\left[T_{go} \leq t \mid T_{go} < T_{stop} + t_d; \theta_{go}; \theta_{stop}\right] \\ &= \frac{1}{p_r(t_d, \theta_{go}, \theta_{stop})} \int_0^t f_{go}(t' \mid \theta_{go})[1 - F_{stop}(t' - t_d \mid \theta_{stop})]\, dt', \end{aligned} \tag{7.15}$$

with θ_{go} and θ_{stop} denoting parameters, or vectors of parameters, for the go and stop-signal distribution, respectively.

A number of different estimation methods for parametric models have been developed. Parameter estimation via maximum likelihood requires the likelihood functions for both go and stop-signal conditions that can be written as follows.

Let $\{t_g\}_{g=1,\ldots,G}$ denote a sample of G response times collected in context *GO*. The log-likelihood function becomes

$$\log L(\theta_{go} \mid \{t_g\}) = \sum_{g=1}^{G} f_{go}(t_g \mid \theta_{go}). \tag{7.16}$$

For context *STOP*, we must distinguish stop-signal responses and inhibitions: let $\{t_r\}_{r=1,\ldots,R}$ denote the signal-respond times for a given SSD $= t_d$. This implies the following log-likelihood function:

$$\log L(\theta_{go}, \theta_{stop} \mid \{t_r\}, t_d) = \sum_{r=1}^{R} f_{go}(t_r \mid \theta_{go})[1 - F_{stop}(t_r - t_d \mid \theta_{stop})]. \tag{7.17}$$

Turning to the inhibitions, let $\{t_i\}_{i=1,\ldots,I}$ denote the successful inhibition (stop-signal) times. Because the t_is are not observable, the likelihood of winning at each possible time point must be considered (by integration). For a given SSD $= t_d$, the log-likelihood function is thus given by (Matzke *et al.*, 2013)

$$\log L(\theta_{go}, \theta_{stop} \mid \{t_i\}, t_d) = \sum_{i=1}^{I} \int_{t_d}^{\infty} \{f_{stop}(t_i - t_d \mid \theta_{stop})[1 - F_{go}(t_i \mid \theta_{go})]\}\, dt_i. \tag{7.18}$$

7.4.1 Exponential Model

We start with the exponential model as an illustrative example permitting closed-form predictions. Several more prominent models will follow, including information about suitable parameter estimation methods.

Let T_{go} and T_{stop} follow exponential distributions; the bivariate cdf is

$$\begin{aligned} H(s,t) &= \mathrm{P}\,[T_{go} \le s] \times \mathrm{P}\,[T_{stop} \le t] \\ &= (1 - \exp[-\lambda_{go}\, s]) \times (1 - \exp[-\lambda_{stop}\, t]), \end{aligned}$$

for all $s, t \ge 0$ with positive real-valued parameters λ_{go} and λ_{stop}. Then

$$\begin{aligned} p_r(t_d) &= \int_0^{\infty} f_{go}(t)[1 - F_{stop}(t - t_d)]\, dt \\ &= \int_0^{t_d} f_{go}(t)\, dt + \int_{t_d}^{\infty} f_{go}(t)\, [1 - F_{stop}(t - t_d)] \\ &= 1 - \frac{\lambda_{stop}}{\lambda_{stop} + \lambda_{go}}\, \exp[-\lambda_{go} t_d]. \end{aligned}$$

The pdf of the signal-response distribution is given, for $t > t_d$, by

$$\begin{aligned} f_{sr}(t \mid t_d) &= f_{go}(t)\,[1 - F_{stop}(t - t_d)]/p_r(t_d) \\ &= \frac{\lambda_{go} \exp[-\lambda_{go} t]\ \exp[-\lambda_{stop}(t - t_d)]}{\left(1 - \dfrac{\lambda_{stop}}{\lambda_{stop} + \lambda_{go}}\ \exp[-\lambda_{go} t_d]\right)} \\ &= \frac{1}{K}(\lambda_{go} + \lambda_{stop})\ \exp[-(\lambda_{go} + \lambda_{stop})(t - t_d)], \end{aligned}$$

with $K = \exp[\lambda_{go}\, t_d](1 + \lambda_{stop}/\lambda_{go}) - \lambda_{stop}/\lambda_{go}$. For $t_d = 0$, we have $K = 1$ and the signal-respond density is identical to an exponential pdf for an independent race between T_{stop} and T_{go}, with parameter $\lambda_{go} + \lambda_{stop}$ and $p_r(t_d) = \lambda_{go}/(\lambda_{go} + \lambda_{stop})$.

For $t \leq t_d$, the density simplifies to

$$\begin{aligned} f_{sr}(t \mid t_d) &= f_{go}(t)/p_r(t_d) \\ &= (\lambda_{stop} + \lambda_{go})\ \exp[-\lambda_{go}(t)]. \end{aligned}$$

Computation of the expected value of signal-response RTs yields

$$\begin{aligned} \mathrm{E}[T_{go} \mid T_{go} < T_{stop} + t_d] &= \int_0^\infty t f_{sr}(t \mid t_d) dt \\ &= \frac{\lambda_{go}\,[1 + (\lambda_{go} + \lambda_{stop}) t_d]}{(\lambda_{go} + \lambda_{stop})\{\exp[\lambda_{go}\, t_d](\lambda_{go} + \lambda_{stop}) - \lambda_{stop}\}}. \end{aligned}$$

In particular, for $t_d = 0$, we obtain $\mathrm{E}[T_{go} \mid T_{go} < T_{stop} + 0] = 1/(\lambda_{go} + \lambda_{stop})$, consistent with the density we mentioned above for this value of the stop-signal delay.

The exponential distribution does not possess a plausible shape as RT distribution, but it is a special case of the *Weibull* distribution that has just one more parameter. The Weibull is often considered to approximate empirical RT distributions, but the Weibull model has not yet been considered for stop-signal modeling, to our knowledge.

7.4.2 Ex-Gaussian Model

This model, explored by Matzke and colleagues (Matzke *et al.*, 2013), relies on the convolution of an exponential and a normal distribution (ex-Gaussian). The ex-Gaussian distribution is described by three parameters: μ and σ the mean and standard deviation of the Gaussian component, and τ the mean of the exponential component.[10] It has a positively skewed unimodal shape with μ and σ reflecting the leading edge and τ the tail of the distribution (see Figure 7.7). It often produces an excellent fit to empirical RT distributions.

10 Note that here τ is the inverse of the λ parameters in the previous model where the exponential mean was $1/\lambda$.

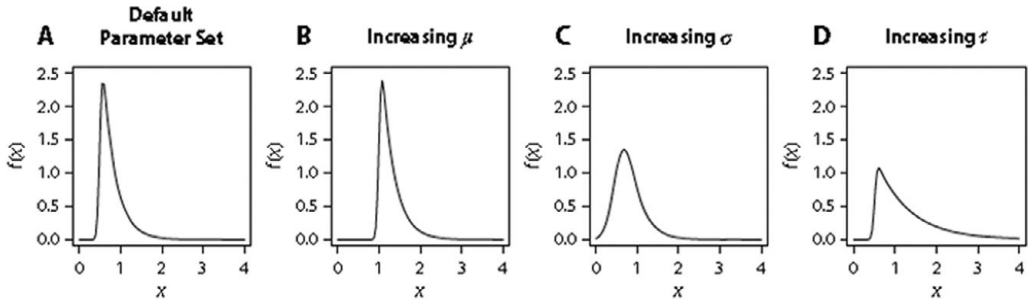


Figure 7.7 *Dependence of ex-Gaussian distributional shape on parameter changes. The parameter sets used to generate the distributions are (A)* $\mu = 0.5$, $\sigma = 0.05$, $\tau = 0.3$ *(default parameter set); (B)* $\mu = 1$, $\sigma = 0.05$, $\tau = 0.3$ *(increasing* μ*); (C)* $\mu = 0.5$, $\sigma = 0.2$, $\tau = 0.3$ *(increasing* σ*); and (D)* $\mu = 0.5$, $\sigma = 0.05$, $\tau = 0.8$ *(increasing* τ*) (from Matzke & Wagenmakers, 2009).*

The pdf of the ex-Gaussian is

$$f(t;\mu,\sigma,\tau) = \frac{1}{\tau}\exp\left[\frac{\mu - t}{\tau} + \frac{\sigma^2}{2\tau^2}\right]\Phi\left[\frac{t-\mu}{\sigma} - \frac{\sigma}{\tau}\right], \tag{7.19}$$

where Φ is the standard normal cdf and $\sigma > 0, \tau > 0$. Moreover, as a sum of two random variables, the expected value equals $\mu+\tau$ and, by stochastic independence of the component distributions, the variance is $\sigma^2 + \tau^2$. Skewness is determined solely by the exponential component and is equal to $2^{1/3}\tau$ (see also Figure 7.7). The ex-Gaussian model has the theoretical defect of predicting negative RTs with positive probability, but this probability can be made arbitrarily small by shifting the distribution to the right.

The ex-Gaussian stop-signal model assumes separate parameter sets for the T_{go} and T_{stop} distributions, $(\mu_{go}, \sigma_{go}, \tau_{go})$ and $(\mu_{stop}, \sigma_{stop}, \tau_{stop})$. Due to the normal component, no closed-form expressions for $F_{sr}(t)$ and $p_r(t_d)$ are available, but simulation is simple by sampling from the two component distributions and adding the values.

Parameter estimation. While model parameter estimates can be obtained via standard maximum likelihood methods, Matzke and colleagues (Matzke *et al.*, 2013) have also developed a Bayesian estimation method to fit the model to both individual and group data.

First, a uniform prior distribution is assumed for the six parameters of the T_{go} and T_{stop} distributions. These priors are informative in the sense that they cover a wide but realistic range of values informed by results from the stop-signal literature (Band *et al.*, 2003). The prior distributions are then updated by the data to yield the posterior distributions, according to Bayes' rule (without marginal likelihood):

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

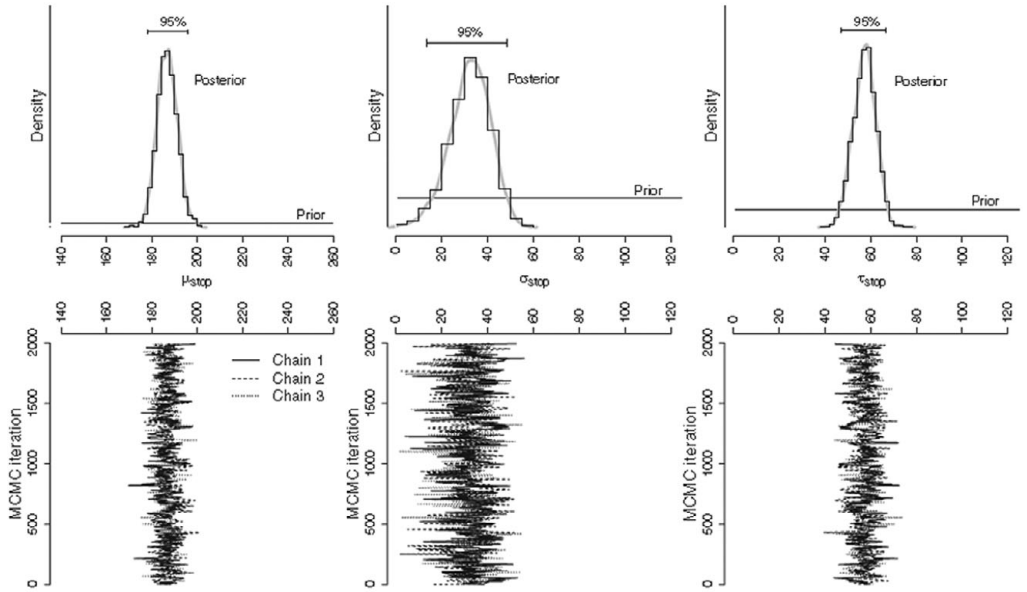


Figure 7.8 *The histograms in the top panel show the posterior distribution of the stop-signal parameters (synthetic data set). The corresponding thick gray lines indicate the fit of a nonparametric density estimator to the posterior samples. The horizontal black lines at the bottom show the prior distribution of the parameters. The horizontal black lines at the top show the 95% Bayesian confidence interval. The solid, dashed, and dotted lines in the bottom panel represent the different sequences of values (i.e., MCMC chains) sampled from the posterior distribution of the parameters (Gibbs sampling) (from Matzke* et al.*, 2013).*

For each parameter, the mean, median, or mode of the posterior distribution is taken as a point estimate of the parameter, while the dispersion of the posterior distribution, quantified by the standard deviation or the percentiles, yields information about the precision of the parameter estimates. The larger the posterior standard deviation, the greater the uncertainty of the estimated parameter. The posterior distribution for each parameter is approximated via *Gibbs sampling* (Geman & Geman, 1984), a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations when direct sampling is difficult (for details, see Matzke *et al.*, 2013). Figure 7.8 illustrates the result for the three parameters of the posterior stop-signal pdf.

The Bayesian parametric approach can also handle group data via hierarchical modeling (Gelman & Hill, 2007). Individual parameters are assumed to be drawn from group-level distributions that specify how the individual parameters are distributed in the population. Given that in stop-signal experiments often relatively few observations per participant are available, the hierarchical approach is especially valuable here. For further details about the estimation procedure and

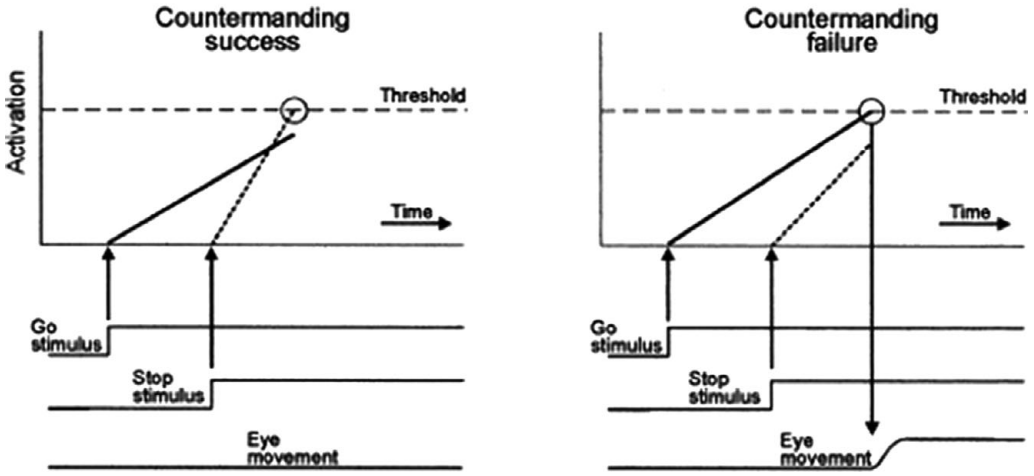


Figure 7.9 *The go process (solid line) and the stop process (dotted line) race independently toward their respective thresholds (dashed horizontal line). The thresholds for both processes coincide only for ease of illustration. In stop trials, the stop process is evoked after the go process has begun. Left panel: The go and stop stimuli each trigger a signal rising linearly toward a threshold; if, as here, the stop process rises so fast that it overtakes the go process and reaches the threshold first, the saccade is successfully inhibited. Right panel: If the go process reaches the threshold first, the saccade fails to be countermanded (from Hanes & Carpenter, 1999).*

accompanying software, we must refer to the original sources (Matzke, 2013; Matzke *et al.*, 2013).

The ex-Gaussian model yields precise information about the unobservable stop-signal times but does not attach a specific substantive meaning to the choice of the distribution. In contrast, the following models motivate their distributional form by certain processing assumptions in the stop-signal task.

7.4.3 Hanes–Carpenter Race Model

The model is based on the linear approach to threshold with ergodic rate (LATER) model purporting to describe the neural mechanism controlling the latency between the appearance of a visual target and the start of a saccadic eye movement to the target (Carpenter & Williams, 1995). Introduced in Hanes and Carpenter (1999), it assumes that the competing go and stop processes rise in a linear fashion to a fixed response threshold. Assuming a fixed response threshold θ, stochastic variability is built into the model by postulating a normally distributed random rate of rise for going and stopping.

The LATER model assumes a linear rise r of the go process to a fixed threshold, starting from an initial activity level s_0, i.e., $s_0 + r \times t = \theta$ (see Figure 7.9).

Assuming r to be the realization of a normally distributed random variable R with mean μ_{go} and variance σ^2_{go}, the above equation leads, after rearrangement, to an expression for the go-process random variable T_{go}:

$$T_{go} = (\theta - s_0)/R.$$

Since the distribution of R is given, the pdf of T_{go} follows as (see Colonius, Özyurt, & Arndt, 2001)

$$f_{go}(t) = \frac{\theta - s_0}{\sigma_{go}\sqrt{2\pi}\, t^2} \exp\left[-\left(\frac{\theta - s_0}{t} \right)^2 / (2\sigma^2_{go}) \right]. \tag{7.20}$$

An analogous pdf is assumed for the stop process T_{stop} with mean μ_{stop} and variance σ^2_{stop} and predictions from the Hanes–Carpenter model are obtained by inserting these distributions into the expression for the signal-respond distribution [Equation (7.15)] and the analogous expression for $p_r(t_d)$. The model has been tested in several studies. Hanes and Carpenter (1999) reported that the model correctly predicted the probability of successful saccade inhibition as a function of the stop-signal delay as well as the signal-respond distributions. Colonius, Özyurt, & Arndt (2001) found results paralleling those of the nonparametric Logan–Cowan model applied to the same data set, and showed that saccade inhibition is more efficient in response to auditory stop signals than visual stop signals.

Parameter estimation has been performed by minimizing sum-of-squares deviations between observed and predicted data using expressions for the pdfs, by maximum likelihood estimation and by Monte Carlo simulations (see, e.g., Colonius, Özyurt, & Arndt, 2001).

7.4.4 Diffusion Race Model Including its Extension to Choice RT

In the Hanes–Carpenter model, stochastic variability is implemented across trials by the random rise of activity in going and stopping, but once started, activation accumulates in a linear deterministic fashion within the trial. In contrast, the *diffusion race model* developed in Logan *et al.* (2014) assumes that both processes are governed by diffusion processes that race against each other until the first one reaches a fixed threshold. The concept of a stochastic diffusion (Wiener) process has arguably become the most important component of modeling response times in a wide variety of tasks (e.g., Busemeyer & Townsend, 1993; Diederich, 1995; Ratcliff, 1978; Smith & Ratcliff, 2009; van Zandt, Colonius, & Proctor, 2000); for details, see Diederich and Mallahi-Karai (2018) and Smith (2000).

The diffusion race model assumes a Wiener diffusion process with drift rate ξ, a starting point of zero activation, and a threshold (absorbing boundary) at z. The first-passage time is given by the inverse Gaussian (or Wald) distribution; for the go-process pdf, we get

$$f_{go}(t) = z(2\pi t^3)^{-0.5} \exp\left[-\frac{1}{2t}(\xi t - z)^2 \right], \tag{7.21}$$

and for the stop process pdf

$$f_{stop}(t) = z(2\pi(t-t_d)^3)^{-0.5} \exp\left[-\frac{1}{2(t-t_d)}(\xi(t-t_d)-z)^2\right] \tag{7.22}$$

for $t > t_d$, and zero otherwise.

The model actually tested in Logan *et al.* (2014) was an extended version of the above, reintroducing across-trial variability by assuming the threshold to be a uniform random variable Z ranging from $z-a$ to $z+a$, with a mean of z and a variance of $a^2/3$. For example, the finishing time of the go process unconditioned over the variable threshold Z results in a go pdf

$$g_{go}(t \mid z, \xi) = (2a)^{-1} \int_{z-a}^{z+a} f_{go}(t \mid z', \xi)\, dz'.$$

The context for this model extension was that the authors were interested in modeling a more general paradigm, where participants' go response was a decision among stimuli from a set A of possible response alternatives. In this paradigm, participants also produce error RTs (choosing the wrong alternative), and it is well known that the diffusion model with constant threshold cannot predict the often observed "fast error" RT distributions (Smith, 2000).

The diffusion race model is an instantiation of what Logan and colleagues call the *general independent race model*. The latter assumes a double race, first, between a set A of possible go responses and second, between the winner of the first race and the stop process. Assuming stochastic independence throughout, this implies for the probability that go response i ($i \in A$) will occur given SSD $= t_d$:

$$\mathrm{P}[\text{response } i \mid t_d] = \int_0^\infty f_i(u) \prod_{j \in A, j \neq i} (1 - F_j(u))\,(1 - F_{stop}(u - t_d))\, du,$$

where $F_j(t)$ and $f_j(t)$ ($j \in A$) are the cdf and pdf, respectively, for go response j. The probability that the stop process wins the race is

$$p_{stop}(t_d) = \int_0^\infty f_{stop}(u - t_d) \prod_{i \in A} (1 - F_i(u))\, du;$$

thus, $p_r(t_d) = 1 - p_{stop}(t_d)$.

For the pdf of RTs conditioned on response i, we get the signal-respond distribution

$$f(t|i, t_d) = \left[f_i(t)(1 - F_{stop}(t)) \prod_{j \in A, j \neq i} (1 - F_j(t)) \right] / p_r(t_d).$$

The pdf for T_{go}, the RT to give some response when no stop signal is present, is

$$f_{go}(t) = \sum_{i \in A} f_i(t) \prod_{j \in A, j \neq i} (1 - F_j(t)).$$

In analogy to Equation (7.9), the signal-respond pdf can be calculated using

$$f_{sr}(t \mid t_d) = \left[\sum_{i \in A} f_i(t) \prod_{j \in A, j \neq i} (1 - F_j(t))(1 - F_{stop}(t - t_d)) \right] / p_r(t_d).$$

This model clearly generalizes the Logan–Cowan race model in covering choice RT paradigms as well. As such, it could be studied further as a nonparametric model, e.g., with an additional assumption of constant SSRT.

However, Logan *et al.* (2014) were specifically interested in issues of processing capacity. For example, do stop and go processes share capacity, or is processing capacity unlimited in the stop-signal paradigm? To answer this question, they systematically varied the number of response alternatives and estimated parameters of the race diffusion model. They hypothesized that, under limited capacity, the rate parameter for the stop process should decrease with the number of alternative responses, just as the rate parameters for the go process do. This is basically what they found using a series of model variants with certain parameters fixed and others free to vary. Moreover, the threshold parameter for the go task increased slightly with the number of alternatives, which is interpreted as subjects adjusting the threshold strategically to compensate for the increased noise.

7.5 Parametric Dependent Race Models

7.5.1 Evidence Against Independence: The Paradox

All models considered up to here were based on assuming both context and stochastic independence. Nevertheless, some recent findings, adding to some earlier ones, have raised serious doubts about the ubiquitous validity of the independence assumptions. A specific independence test is to check that mean signal-respond RTs are monotonically increasing with stop-signal delay and that corresponding distribution functions are ordered accordingly (see Figure 7.10, left panel). In earlier work, we have found some violations of this ordering at short SSDs (e.g., Colonius, Özyurt, & Arndt, 2001; Özyurt, Colonius, & Arndt, 2003; see Figure 7.10, right panel) but evidence remained weak because, typically, observations are sparse at short SSDs. Moreover, Band, van der Molen, and Logan (2003), investigating the consequences of violations of both context and stochastic independence on stop-signal processing estimates via simulation, found severe bias effects on SSRT estimates under some conditions. Recently, in a large-scale survey analyzing 14 experimental studies, Bissett *et al.* (2021) found serious violations of context independence specifically at short SSDs (i.e., less than 200 ms).

Such violations are commonly interpreted as refuting context independence, but it seems difficult to tell apart violations of stochastic independence from violations of context independence by experimental tests of behavior. Thus, violations of the former, in addition to or in place of violations of context independence cannot be ruled out.

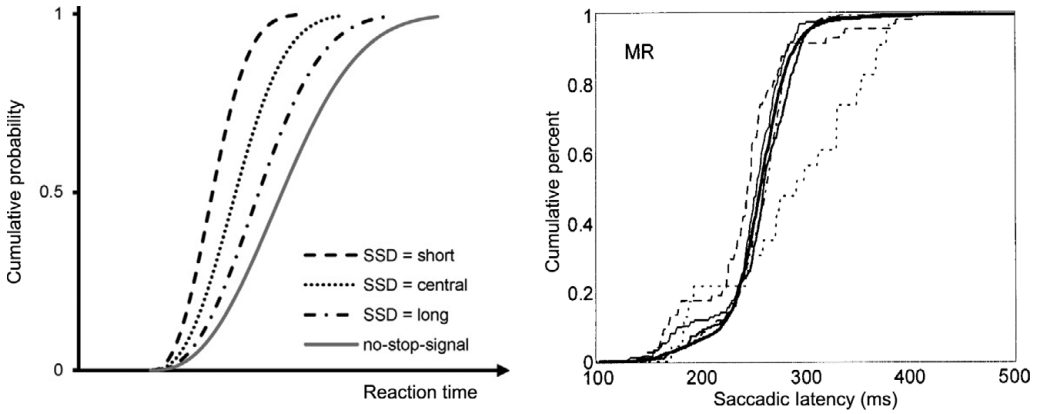


Figure 7.10 *Distribution function (cdf) prediction of IND model. Left panel: Prediction for signal-respond cdfs ordered by SSD size (from Verbruggen & Logan, 2009). Right panel: Observed violation for short SSD = 90 ms (from Colonius, Özyurt, & Arndt, 2001). In both panels, the solid line represents the go-signal RT distribution.*

Strong evidence against independence comes from seminal findings on the neural underpinnings of response inhibition, and the main impetus for developing race models with dependency arguably comes from these investigations. Studies in the frontal eye fields (FEFs) and superior colliculus (SC) of macaque monkeys performing a countermanding task with saccadic eye movements have shown that the neural correlates of go and stop processes produce eye movement behavior through a network of interacting gaze-shifting and gaze-holding neurons (Brown *et al.*, 2008; Hanes, Patterson, & Schall, 1998; Hanes & Schall, 1995; Middlebrooks *et al.*, 2020; Paré & Hanes, 2003). Specifically, Hanes and colleagues (Hanes & Schall, 1995) showed, first, that macaque monkey behavior in saccade countermanding corresponded to that of human performance in manual stop-signal tasks consistent with the independent model. Then, recording from FEFs they isolated neurons involved in gaze shifting and gaze holding that represent a larger circuit of such neurons that extends from the cortex through the basal ganglia and SC to the brainstem (see Figure 7.11).

The question thus arises: How can *interacting* circuits of mutually inhibitory neurons instantiate stop and go processes with (context or stochastically) independent finishing times? Although it can be argued that behavioral and neural data provide a description on different levels of processing (see Section 7.5.3 below), this discrepancy has widely been perceived as a paradox (Boucher *et al.*, 2007; Matzke, Verbruggen, & Logan, 2018; Schall & Godlove, 2012; Schall, Palmeri, & Logan, 2017).

In an effort to resolve the paradox, a number of neurally inspired, computationally explicit models have been proposed that will be considered here and in the following section. In Section 7.7, we will present some further behaviorally oriented approaches based on recent concepts of statistical dependence.

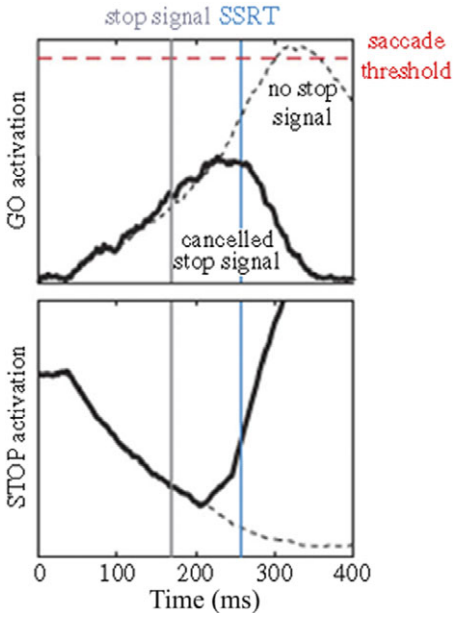


Figure 7.11 *Schematic diagram: Activation of the GO unit (upper) and the STOP unit (lower) for trials with no stop signal (dashed lines) and trials with a stop signal that successfully canceled the saccade (solid lines). Saccades are produced when inhibition of the STOP unit is released and the activation of a GO unit reaches a threshold (topmost dashed line). In response to the stop signal (left vertical line), the STOP unit becomes active, interrupting the accumulation of GO unit activation. This interruption occurs immediately before the stop-signal reaction time (SSRT) (right vertical line), a measure of STOP process duration derived from the independent race model (from Schall, Palmeri, & Logan, 2017).*

7.5.2 Interactive Race Model

Boucher and colleagues (Boucher *et al.*, 2007) developed a relatively simple neural network model, the *interactive race model*, consisting of a go (or move) and a stop (or fixation) unit that accumulate stochastic evidence and race toward a common threshold (arbitrarily set to one). Whichever unit reaches the threshold first determines whether a stop signal trial is signal-inhibit or signal-respond.

The approach, based on a version of the *leaky, competing accumulator model* (Bogacz *et al.*, 2006; Usher & McClelland, 2001), is defined by two stochastic differential equations:

$$da_{go}(t) = \frac{dt}{\tau}\left[\mu_{go} - k\,a_{go}(t) - \beta_{stop}\,a_{stop}(t)\right] + \sqrt{\frac{dt}{\tau}}\,\xi_{go}; \tag{7.23}$$

$$da_{stop}(t) = \frac{dt}{\tau}\left[\mu_{stop} - k\,a_{stop}(t) - \beta_{go}\,a_{go}(t)\right] + \sqrt{\frac{dt}{\tau}}\,\xi_{stop}. \tag{7.24}$$

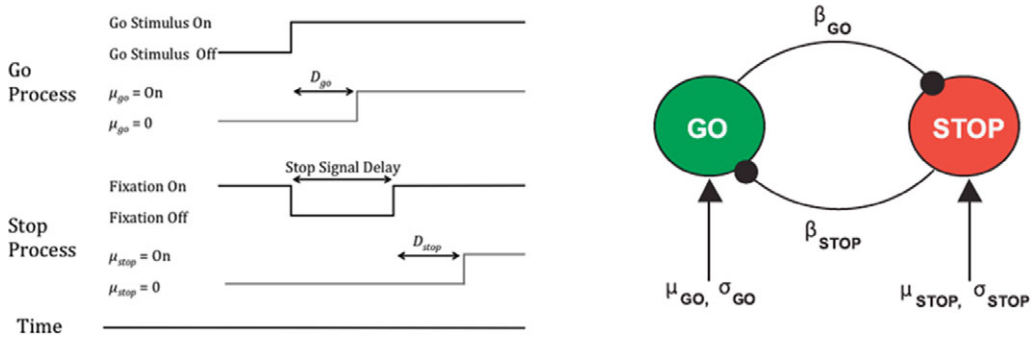


Figure 7.12 *Interactive race model. Left panel: Timing of events (including afferent delays D_{go} and D_{stop}) (adapted after Logan* et al.*, 2015). Right panel: Go and stop unit interaction (with leakage parameter k dropped) (adapted after Boucher* et al.*, 2007).*

These equations describe the change in activation of the go and stop units, $a_{go}(t)$ and $a_{stop}(t)$, within (an infinitely small) time step dt. Parameters μ_{go} and μ_{stop} denote mean growth rates (*drift rates*) for the go and stop unit, respectively. The leakage parameter, k, prevents the activation from increasing without bound. Interaction between the units is controlled by the inhibition parameters β_{go} and β_{stop} (see Figure 7.12, right panel). The amount of mutual inhibition depends on the instantaneous activation levels, $a_{go}(t)$ and $a_{stop}(t)$, causing a unit with a low activation to have a small inhibitory effect on the other unit. Finally, ξ_{go} and ξ_{stop} are Gaussian noise terms with mean zero and variance σ^2_{go} and σ^2_{stop}, respectively.

Other parameters in the model capture the non-decision time stages of processing. Stimulus encoding that occurs before go unit and stop unit activation was instantiated is represented as constant delay: D_{go} denotes afferent processing time after the go stimulus is presented, D_{stop} is the latent time after SSD and before the stop unit begins to inhibit the go unit (see Figure 7.12, left panel). Boucher *et al.* (2007) studied simultaneously recorded behavioral and neural data from two monkeys performing the saccadic stop-signal task (Hanes, Patterson, & Schall, 1998). Because the above model equations do not possess closed-form solutions, they simulated the model searching for optimal parameter values to minimize deviations of predictions from the data. In order to fit neurophysiological data, they first had to decide which parts of the neural populations should correspond to the stop and go units of the interactive model. They noticed:

> The stop-signal task is ideal for investigating the neural control of movement initiation because it specifies the criteria a neuron must meet to be identified as contributing to controlling saccade initiation. First, the activity in trials when a saccade is made (no-stop-signal or signal-respond trials) must be different from that in trials when no saccade is made (signal-inhibit trials). Second, in stop-signal trials, the activity should begin along the trajectory that would lead to saccade initiation, but on presentation of the stop signal, the activity must be modulated

> away from that trajectory, and this modulation must occur within the SSRT. Neurons with movement-related and fixation-related activity in frontal eye field and superior colliculus satisfy both of these requirements. (Boucher *et al.*, 2007, p. 380)[11]

Second, go-unit activation was compared with movement neuron activity and stop-unit activation was compared with fixation neuron activity. Specifically, for both neurons and model units, activation on signal-inhibit or signal-respond trials was compared with the activity of a subset of latency-matched no-stop-signal trials. No-stop-signal trials with response time longer than SSD + SSRT were compared with signal-inhibit trials, because according to the race model, the saccade would have been inhibited had the stop signal been presented. No-stop-signal trials with response time shorter than SSD + SSRT were compared with signal-respond trials because, according to the race model, the saccade would have been initiated even if the stop signal had occurred. *Cancel time* was defined as the time at which activation on signal-inhibit trials significantly diverged from the activation on no-stop-signal trials relative to SSRT[12] (for further details, refer to Boucher *et al.*, 2007, p. 386).

In probing the model, Boucher and colleagues first evaluated the ability of the independent race model to account for the observed data. Setting inhibition parameters β_{go} and β_{stop} to zero turns the model into a stochastically independent version, and this resulted in good fits to the behavioral data (inhibition functions, go RT, and signal-respond RT distributions). However, since it has no mechanism to shut off the go process so that it does not reach the threshold on signal-inhibit trials when the stop process wins, it could not account for the neural data. On the other hand, letting parameters free to vary and utilizing some additional model simulations to estimate go and stop-signal cancel times, the authors showed that the interactive model can be fitted simultaneously to both neural and behavioral data. Moreover, by constraining the model parameters in different ways, it turned out that a good model fit depended on two restrictions: (i) activation of the stop unit has to be delayed for a substantial amount of time after the stop signal occurs, i.e., D_{stop} must be rather large (50–70 ms) and (ii) the stop unit must inhibit the go unit much more than vice versa, i.e., β_{stop} has to be much larger (i.e., by an order of magnitude) than β_{go}.

What are the consequences of these findings for the interpretation of SSRT, as measured in the Logan–Cowan independent race model? The parameterized interactive race model implies an additive partition of SSRT. First, the stop-signal

11 As the authors point out, determining the quantitative details is a rather subtle task. First, neural activation functions derived from spike trains are converted to spike density functions, as described in Hanes, Patterson, and Schall (1998). Although a neural population with a specific function should respond in generally the same way, each neuron may have some idiosyncrasies. Thus, before averaging across neurons, they first had to normalize the spike density function of each neuron by dividing its activity by the peak firing rate in the interval from 20 ms before to 50 ms after saccade initiation on no-stop-signal trials.

12 Cancel time is important in neuroscience because it is an essential criterion for determining whether modulation of neural activity happens early enough to participate in response inhibition (Logan *et al.*, 2015, p. 123).

encoding time, D_{stop}, was between 51 and 67 ms with a small standard deviation (10–20 ms). Second, the interval from SSD+D_{stop} until cancel time (interruption of go-unit accumulation), called $stop_{interrupt}$, is only about 22 ms, effectively instantaneous. Adding the ballistic interval preceding initiation of the movement (Logan & Cowan, 1984), denoted as $go_{ballistic}$ (about 10 ms), results in the following SSRT decomposition:

$$\text{SSRT} = D_{stop} + stop_{interrupt} + go_{ballistic}. \tag{7.25}$$

With SSRT estimates from behavioral data in the range of 80–95 ms, this equation means that most of SSRT is occupied by D_{stop}, during which the go unit is not affected by the stop unit. Boucher *et al.* (2007) conclude that stopping is a two-stage process consisting of a (relatively long) encoding stage with no interaction and a brief interruption stage during which response preparation is inhibited. This model has been postulated to be a resolution of the above-mentioned paradox of an independent race at the level of RTs and mutual inhibition at the level of neural activation between gaze-holding and gaze-shifting units (see also Schall *et al.*, 2017).

7.5.3 Linking Propositions

The general race model and most of its subclasses do not make a commitment to the underlying computational or neural processes that generate the processing times T_{go} and T_{stop}. The interactive race model, however, has been developed with the aim of connecting go and stop-signal processing to the underlying physiology. Given the good understanding of how saccade production is controlled by a circuit of neurons extending from the cortex through the basal ganglia and superior colliculus to the brainstem, the model links the go unit to movement-related neurons and the stop unit to fixation-related neurons in frontal eye fields and superior colliculus (Boucher *et al.*, 2007).

Such *linking propositions* specifying the nature of the mapping between particular cognitive states and neural states have a long history (e.g., Teller, 1984) and they have recently become more popular under the heading of *model-based cognitive neuroscience* (e.g., Forstmann & Wagenmakers, 2015). One motivation for developing linking propositions is the hope to solve the general problem of non-identifiability and model mimicry (see Jones & Dzhafarov, 2014), that exists for behavioral models of choice RT, by identifying the underlying neurophysiology. In the context of the stop-signal task, Schall and colleagues have thoroughly investigated linking propositions between processing times (T_{go}, T_{stop}) and single-neuron discharges in the frontal eye field, superior colliculus, and ocular motor neurons leading to the interactive race model of Section 7.5.2 and related models (Schall, 2004, 2019; Schall & Godlove, 2012). Unfortunately, as recently described in Schall (2019), finding a one-to-one mapping between parameters of neural activity and those describing abstract stochastic accumulators (like in race models) seems out of reach at the moment (see also Schall & Paré, 2021).

7.6 Related (Non-race) Models

7.6.1 Blocked-Input Model

Starting from the interactive model, Logan and colleagues (Logan *et al.*, 2015) suggested an alternative view on how stopping occurs: the stop unit does not directly inhibit growth of activation of the go unit; rather, the stop signal activates a top-down process outside the gaze control network that, once reaching a threshold early enough, blocks input to the go unit. Within the dynamics of the interactive model, this means setting the go drift rate to zero so that it will not reach its threshold.

The authors defined the units more neutrally as fixation (fix) and movement (move) units because they linked them to gaze-holding and gaze-shifting neurons in a general network, extending from the cerebral cortex to the brain stem and being in active balance already at the start of a trial. Modeling steady-state fixation activity implied that eye movements can only occur if activation in the move unit (μ_{move}) and inhibition from the move unit to the fix unit (β_{move}) are large enough to overcome steady-state activation in the fix unit, and if simultaneously inhibition from the fix unit to the move unit (β_{fix}) is not large enough to suppress move activation entirely. For the monkey FEF data from Hanes, Patterson, and Schall (1998), these constraints led to equivalent predictions of physiological data for the interactive and the blocked-input model, but the latter model provided a better account of the behavioral data.

By letting certain model parameters vary freely and keeping others fixed, Logan *et al.* (2015) compared fits of different versions of the interactive and blocked-input models. Although these models differed strongly with respect to the temporal dynamics of inhibition, they did not show substantial differences in goodness of fit. The authors refer to this result as an instance of "model mimicry" of blocking and inhibiting which can only partially be resolved by considering neurophysiological data.

7.6.2 DINASAUR Model

A recent neural network model by Bompas and colleagues (Bompas, Campbell, & Sumner, 2020) tackles the problem of modeling rapid saccadic countermanding from a different background. Their model had originally been developed for the well-known phenomenon of *saccadic inhibition* (SI) (Bompas & Sumner, 2009; Reingold & Stampe, 2002; Walker & Benson, 2013).

SI occurs in a paradigm that is, or can be made to be, identical to the stop-signal task in all aspects except for the instruction: instead of inhibiting the response upon appearance of a (stop) signal, the participant is instructed to just ignore it and perform the saccade to the target stimulus. The SI effect is manifest as a decrease in the number of saccades observed shortly after (distractor) stimulus onset, compared with baseline conditions (with no signal), with a maximum inhibitory influence occurring around 70–90 ms later (see Figure 7.13).

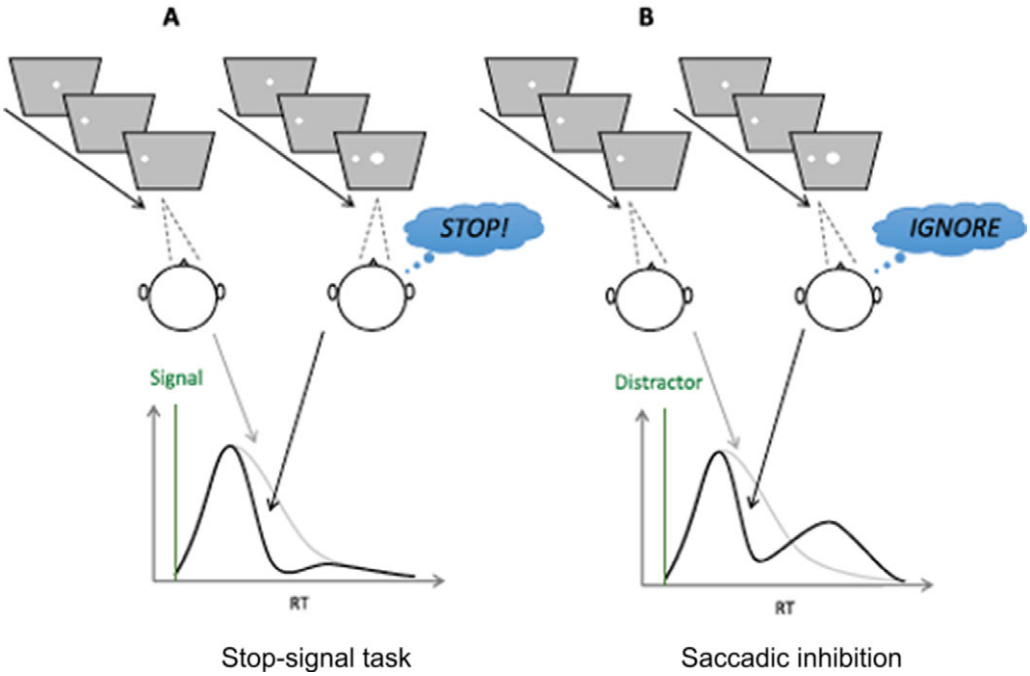


Figure 7.13 *Saccadic stop-signal task (panel A) and saccadic inhibition (panel B) paradigms. Both paradigms involve a stimulus jump from center to periphery, sometimes followed by the onset of a central signal (right subpanels above, black lines below), sometimes not (left subpanels, gray lines). The signal onset time is indicated by the vertical lines and the delay between the target jump and the signal is referred to as the stimulus onset asynchrony (SOA). The two tasks differ in the instruction associated with the signal onset: withhold the saccade vs. ignore the signal and perform the saccade. Panel A: Instructions to stop remove slower responses from the RT distribution, but fast responses escape. Panel B: The same visual events associated with an ignore instruction typically produce a dip in the latency distribution, where saccades are delayed and subsequently recover, so that the total number of saccades is about the same between signal present and no-signal distributions (adapted after Bompas, Campbell, & Sumner, 2020).*

The authors start from the observation that in the stop-signal paradigm, as in SI, fixation and movement neurons receive inputs tightly tied to the visual stimuli (targets and stop signals), with onsets and offsets leading to step changes some 35–50 ms later, preceding inputs from control neurons whose role is to cancel the action plan (Bompas, Campbell, & Sumner, 2020, p. 528). The first part of rapid saccadic countermanding is initially entirely automatic, with slower, top-down endogenous signals built on top of rapid automatic disruption. Their model refers to an approach originally developed by Trappenberg and colleagues (Trappenberg *et al.*, 2001) describing the dynamics of saccadic decision with basic characteristics of exogenous and endogenous neural signals and lateral inhibition in the intermediate layers of the superior colliculus (SC).

A specific instantiation of Bompas *et al.*'s model, called 200N-DINASAUR, possesses $n = 200$ nodes representing the horizontal dimension of the visual field, and the average spiking rate A_i of neuron i is a logistic function of its internal state u_i:

$$A_i(t) = 1/(1 + \exp[-\beta\, u_i(t)]).$$

Similar to leaky competing accumulators models, the dynamics of $u_i(t)$ across time depends on normally distributed noise and two types of input received, either external to the map (endogenous or exogenous) or internal via lateral connections (cf. Trappenberg *et al.*, 2001):

$$\tau\, \frac{du_i}{dt} = -k\, u_i(t) + \frac{1}{n} \sum_j \omega_{ij}\, A_j(t) + I_i^{exo}(t) + I_i^{endo}(t) + N(0, \eta). \tag{7.26}$$

The authors emphasize the distinction between visual events triggering exogenous inputs (i.e., transients tied to visual changes: targets, distractors, or stop signals) not affected by instructions, and endogenous signals (i.e., later, sustained, and linked to the instructions) (Bompas, Campbell, & Sumner, 2020, p. 529). Endogenous inputs vary as step functions, while exogenous inputs are transient, reaching their maximal amplitude (a_{exo}) at $t = t_{onset} + \delta_{vis}$, and then decreasing exponentially as a function of time, according to the following equation:

$$\tau_{on}\, \frac{dI_i^{exo}}{dt} = -I_i^{exo}(t) + a_{exo}.$$

Following the Trappenberg *et al.* model, all inputs have Gaussian spatial profiles (with standard deviation σ). They are maximal at the targeted nodes but also affect nearby nodes. Lateral connections show a Gaussian spatial profile that changes from positive (excitation) at short distance to negative (inhibition) at longer distance according to connection weights ω_{ij} (see Bompas, Campbell, & Sumner, 2020, p. 530).

In the no-signal condition, a single exogenous (visual) transient onset occurs δ_{vis} after target onset, shortly followed by a switch of endogenous support from fixation to target δ_{endo} after target onset. The signal-ignore condition differs from the no-signal condition solely by the presence of a second visual transient, triggered by the signal appearing. When generalizing the model from signal-ignore to signal-stop conditions, only the endogenous input should differ because the visual display is identical and only the instructions differ. As in the blocked-input model, the endogenous input to the target is switched off (blocked) δ_{endo} after the stop-signal, while the endogenous input to the fixation is switched on again.

Bompas, Campbell, and Sumner (2020) validate the DINASAUR model in several steps via both simulation and experiment. While the model features a large number of parameters (up to 16), by taking all but two of the parameters from the model fit for the SI paradigm in Bompas and Sumner (2009), their simulation was

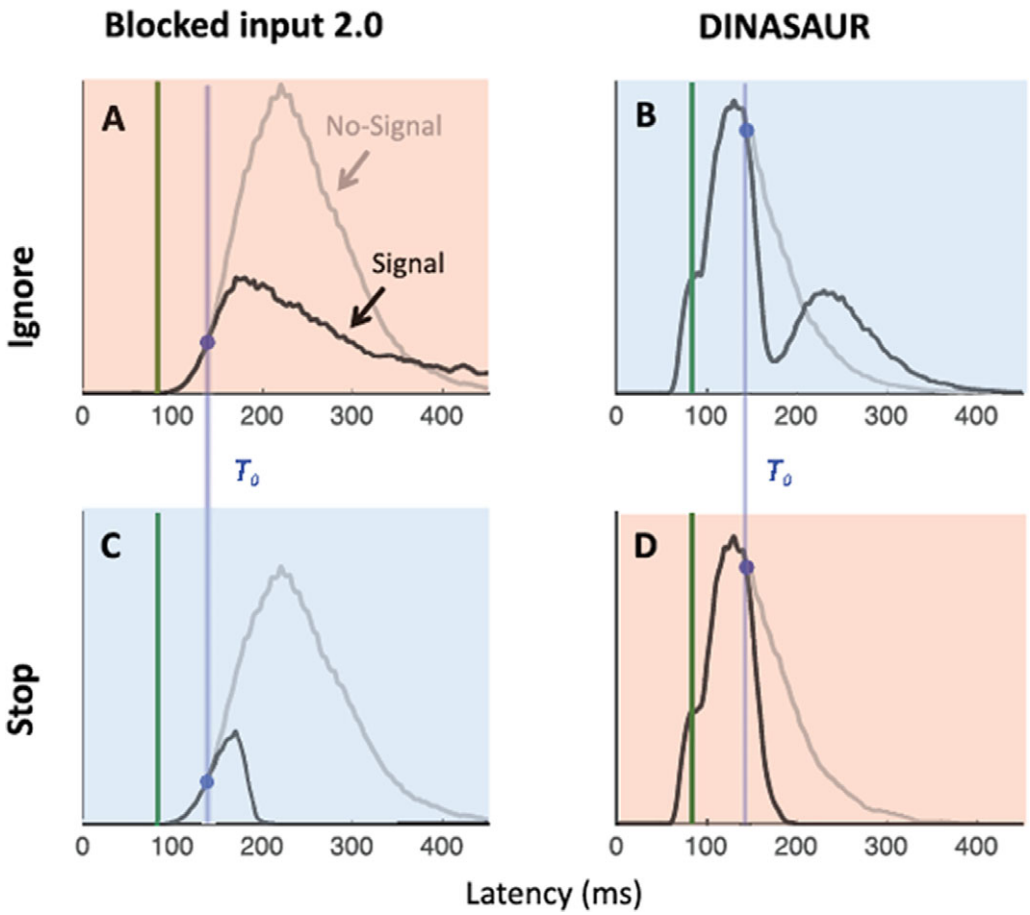


Figure 7.14 *Simulated RT distributions for blocked input (left panels) and 200N-DINASAUR (right panels) model for ignore-signal (upper panels) and stop-signal condition (lower panels). The DINASAUR model (with blocked input for stopping) captures well the typical pattern of results obtained in both paradigms. Blocked input 2.0 (with adding automatic fixation activity for ignore conditions) is not able to produce the sharp dips expected from the saccadic inhibition literature. Both models predict a perfect alignment across instructions of the time when the signal RT distribution (black) departs from the no-signal RT distribution (gray), indicated by the dots (T_0) and highlighted by the long vertical bars (adapted after Bompas, Campbell, & Sumner, 2020).*

able to reproduce well the typical pattern of results obtained in both paradigms[13] (see Figure 7.14).

The model makes two important predictions. Work on SI by Bompas and Sumner (2011) had indicated that dip onset, the time point T_0 where latency

13 Bompas, Campbell, and Sumner (2020) compare their model with the blocked-input model in much more detail, but we do not go into this here.

distributions diverge, matches the sum of sensory delay δ_{vis} and motor output delay δ_{out} so that $T_0 - SOA$ reflects non-decision time. Moreover, following Boucher *et al.* (2007), a large portion of SSRT is devoted to non-decision time (the independent processing part, followed by rapid and late inhibition). Thus, Bompas *et al*. argue that SSRT "... likely behaves like T_0, and therefore we expect the early part of the interference from stop-signals and distractors should be very similar in saccadic inhibition and countermanding" (Bompas, Campbell, & Sumner, 2020, p. 536). Therefore, the first strong prediction of DINASAUR is that the time point at which the RT distribution diverges from the no-signal distribution should be the same under the ignore-signal and the stop-signal instruction (see point T_0 in Figure 7.14, top and bottom right panels).

The second prediction follows from separating exogenous (visual) delay δ_{vis} from endogenous delay δ_{endo}, and from parsimoniously assuming the latter value to be the same in all phases: (a) endogenous support for the target following target onset; (b) the removal of endogenous support for fixation following target onset; (c) the removal of endogenous support for the target following the signal under the stop instruction; and (d) endogenous support returning to fixation following the stop instruction. The prediction then is that extracting these parameters from the no-signal and signal-ignore conditions permits predicting stopping behavior without the need for additional top-down countermanding parameters.

Bompas, Campbell, and Sumner (2020) found support in three experiments geared toward probing these predictions, but only after adding two amendments to improve fits to the no-signal distribution. The first is to introduce a holding period in order to account for the participants' strategic slowing down in the stop task (*proactive inhibition*). Second, in order to predict "late errors" in the stop-signal condition, they had to add a parameter for the probability of not following the stop instruction. This corresponds to the probability of "trigger failures" (see, e.g., Band, van der Molen, & Logan, 2003 and Section 7.8.2 below).

7.6.3 Diffusion-Stop Model

This model does not implement a race concept either and is related to the blocked-input model closely enough to be mentioned here. In an unpublished paper (Colonius & Diederich, 2001/2021), we address the paradox mentioned at the start of this section by suggesting a diffusion model approach based on Diederich (1997).

The diffusion-stop model assumes a variable growth to a fixed threshold. Rather than claiming separate growths of go and stop-signal-related activities, it assumes a single diffusion process unfolding over time between two fixed criterion thresholds. The onset of the go signal triggers a growth process represented by a stochastic trajectory drifting toward the upper boundary, threshold (θ_{go}). In the absence of a stop signal, the average trajectory (indicated by the line in Figure 7.15, left panel) has a positive slope resulting in mean saccadic response time determined by the time point corresponding to the crossing of the go threshold.

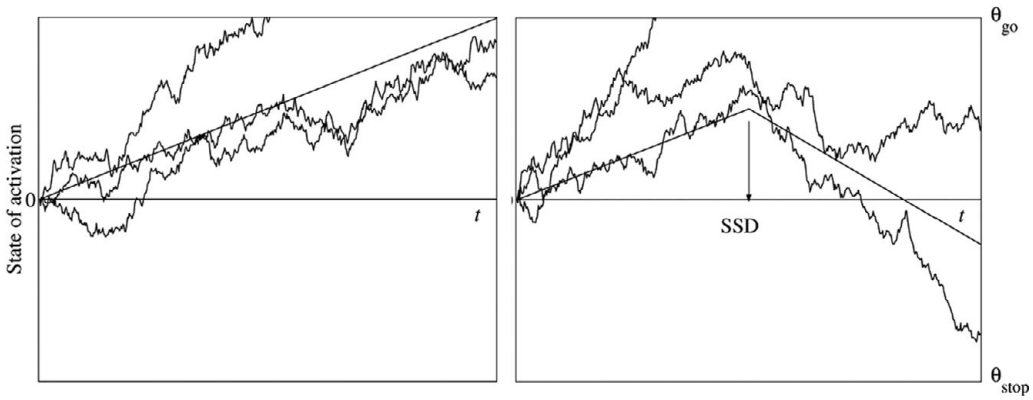


Figure 7.15 *Three hypothetical trajectories in the activation space simulated by the diffusion-stop model. Left: In go trials, the drift rate is constant. A saccade is initiated when the trajectory crosses the upper threshold θ_{go} for the first time. The line presents the average trajectory. Right: In stop-signal trials, the drift rate switches with presentation of a stop signal at SSD. The saccade is inhibited with certainty once the lower threshold θ_{stop} has been crossed the first time. The average trajectory switches slope from positive to negative at SSD.*

On the other hand, crossing the stop threshold (θ_{stop}) results in a permanent cancellation of the planned movement to the go signal. Figure 7.15 illustrates this mechanism. Presentation of a stop signal at point SSD after the go signal shifts the slope of the linear drift to a negative value. Trajectories that have not yet crossed the upper boundary will then tend in the direction of stop criterion θ_{stop}. Due to stochastic variability, however, individual trajectories may still cross the upper boundary, resulting in a response in spite of the stop signal. Note that, like the race model, the diffusion model does not predict different rates of rise in activity for responses in non-canceled trials and in latency-matched no-stop-signal trials. Moreover, consistent with empirical data, the later the stop signal is presented, the less likely a successful inhibition of the saccade becomes.

Specifically, the growth process in the diffusion-stop model is represented by a standard Brownian motion (or Wiener) process $A(t)$ with drift rate $\mu(t)$ and two absorbing barriers θ_{go} and θ_{stop}. The process is *time-inhomogeneous* because the drift rate changes with the occurrence of the stop signal at $t = \text{SSD}$:

$$\mu(t) = \begin{cases} \mu_{go}, & \text{if } t \leq \text{SSD} \\ \mu_{stop}, & \text{if } t > \text{SSD} \end{cases}.$$

The first-passage times are defined as

$$T_{go} = \inf\{t : A(t) \geq \theta_{go} \text{ and } A(\tau) \geq \theta_{stop} \text{ for all } \tau < t\} \text{ and}$$
$$T_{stop} = \inf\{t : A(t) \leq \theta_{stop} \text{ and } A(\tau) \leq \theta_{go} \text{ for all } \tau < t\},$$

with $\theta_{stop} < A(0) < \theta_{go}$.

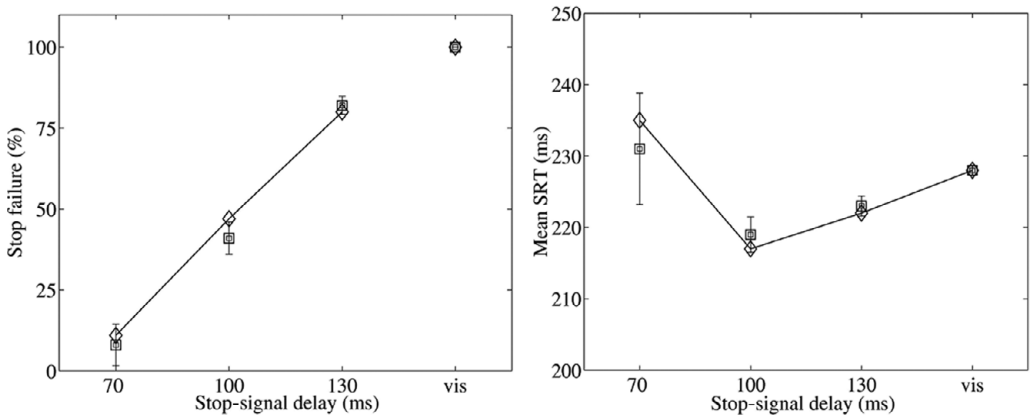


Figure 7.16 *Observed data (squares) and predictions (diamonds) from diffusion-stop model for one subject. Left: Stop failure probability as a function of stop-signal delay. Rightmost point refers to no-stop-signal condition. Right: Mean saccadic response time as a function of stop-signal delay.*

The model was fit to the data of one subject reported in Colonius, Özyurt, & Arndt (2001) using the finite-state Markov chain approximation of the diffusion process (Diederich, 1997). The observable saccadic reaction time was taken as $SRT = T_{go} + c$ (with c a sensorimotor constant) and observable inhibition probability as $\mathrm{P}(T_{go} < T_{stop} + \mathrm{SSD})$. Assuming no bias, $A(0)$ was set to zero. Estimated parameters are the drift rate values μ_{go} and μ_{stop}, the distance between the go and the stop threshold $\theta_{go} - \theta_{stop}$, and constant c. The model fit (seven data points and four parameters) is depicted in Figure 7.16. Note that the diffusion-stop model, in contrast to independent race models, is able to account for the nonmonotonic relation between mean response time and stop-signal delay.

Measuring the speed of the stop process differs strongly from race models. We define *inhibition time* (IT) as the interval from presentation of the stop signal until the trajectory reaches the stop criterion (lower bound). Thus, IT depends on the momentary level of activity towards the go threshold (represented by the trajectory location) at the time the stop signal is presented. It implies that the average time to cancel a saccade increases with stop signal delay. For example, estimates of IT were 530, 570, and 580 ms for stop signal delays of 70, 100, and 130 ms, respectively. Even if one subtracts some 30 ms for the latency of the response to the stop signal, the resulting estimates are one-half order of magnitude larger than the estimates for SSRT under the race model (100 ms in this case). This discrepancy reflects an important difference between the diffusion-stop and the race model: while both IT and SSRT are initiated by the presentation of a stop signal, termination of IT in the diffusion-stop model indicates that inhibition of the saccade has become certain, whereas termination of SSRT in the race model means that stop-signal processing is finished, but actual inhibition of the saccade only occurs if go-signal processing has not been terminated earlier. Estimates for IT of

about 500 ms in the diffusion-stop model may appear implausible, but it should be noted that this includes the time to suppress the go-signal activity completely. If, for example, the go and stop signal are presented nearly simultaneously, resulting in a very high probability of successful inhibition, then estimates for IT can go down strongly, depending on the relative values of the drift parameters.

Subjects have considerable leeway in performing the countermanding task (proactive inhibition). In the diffusion-stop model, a bias in favoring either stopping performance or response speed is easily accounted for by letting a trajectory in the activation space start from a level closer to the stop criterion or to the go criterion, respectively. Introducing this bias parameter also allows the model to predict sequential effects like a higher probability of canceling a saccade if the movement had failed to be canceled on the previous trial (see Section 7.8).

7.7 Semi-parametric Race Models

The assumption of context independence is fundamental to the general race model (Section 7.3.1). In addition, stochastic independence has been assumed in all race models discussed so far, with the exception of the interactive race model (Section 7.5.2). Given that this latter model is fully parameterized, one may wonder whether other race models with stochastically dependent "races" can be developed without making strong assumptions about the distributions of T_{go} and T_{stop}.

7.7.1 The Role of Copulas

It turns out that the concept of a *copula* is a natural tool to investigate such dependent race models. Briefly, a copula is a function that specifies how a multivariate distribution is related to its one-dimensional marginal distributions.[14] For stop-signal modeling, this means that the bivariate distribution H can be written as[15]

$$H(t,s) = \mathrm{P}\left[T_{stop} \le t, T_{go} \le s\right] = \mathrm{C}(F_{stop}(t), F_{go}(s)), \tag{7.27}$$

where C is a bivariate copula that is determined uniquely assuming continuous marginal distributions. Note that a copula specifies the dependency structure without the need to commit to a given distributional family for the marginals, here F_{go} and F_{stop}. For example, letting $u = F_{stop}(t)$ and $v = F_{go}(s)$, copula

$$\mathrm{C}_{IND}(u,v) \equiv u\,v$$

defines stochastically independent race models. Because of the generality of the copula definition, the class of race models based on copulas obviously encompasses all race models with specified marginal distributions.

14 For precise definitions of, and an introduction to, copulas refer to Durante and Sempi (2016), Joe (2015), Nelsen (2006); for an introduction in psychological contexts, see Colonius (2016).

15 Note that in Section 7.7 (only) we write H with the order of marginals switched.

Example: Farlie–Gumbel–Morgenstern Copula. The Farlie–Gumbel–Morgenstern (FGM) copula is defined as

$$C_{FGM}(u,v) = u\,v[1+\delta(1-u)\,(1-v)] \tag{7.28}$$

with parameter δ a real-valued constant. It defines a stochastically dependent *semi-parametric* race model with bivariate distribution function

$$\begin{aligned} H_{FGM}(t,s) &= C_{FGM}(F_{stop}(t), F_{go}(s)) \\ &= F_{stop}(t)F_{go}(s)[1+\delta(1-F_{stop}(t))(1-F_{go}(s))], \end{aligned} \tag{7.29}$$

with parameter δ determining the strength of dependence between T_{go} and T_{stop}. Setting $\delta = 0$ corresponds to the independent race model, negative and positive values of δ to negative or positive dependent models, respectively. It is known that the FGM copula only allows for moderate levels of dependence (e.g., Kendall's tau, $\tau \in [-2/9, 2/9]$).[16]

By inserting specific marginal distributions into a copula, fully parameterized models can be created. For example, with ex-Gaussian marginals with parameters μ and σ for the Gaussian and λ for the exponential component, this results in the ex-Gaussian version of the FGM copula:

$$H_{FGM}(t,s) = F_{stop}(t;\boldsymbol{\theta}_{stop})\,F_{go}(s;\boldsymbol{\theta}_{go})[1+\delta(1-F_{stop}(t;\boldsymbol{\theta}_{stop}))(1-F_{go}(s;\boldsymbol{\theta}_{go}))],$$

where $\boldsymbol{\theta}_{stop} = (\mu_{stop}, \sigma_{stop}, \lambda_{stop})$ and $\boldsymbol{\theta}_{go} = (\mu_{go}, \sigma_{go}, \lambda_{go})$ are parameter vectors, adding up to a total of seven model parameters including δ.

7.7.2 Equivalence with Dependent Censoring

Many alternative copula families with relatively simple dependency structures exist and could be investigated. The specific challenge for stop-signal race models is, of course, that F_{stop} is unobservable. Fortunately, it turns out that the problem of determining the distribution of non-observable stopping time T_{stop} in the race model is formally equivalent to a problem studied in actuarial science concerned with the *time of failure* of some entity (human, machine, etc.). Recall that *censoring* is a condition in which the failure time is only partially known. For example, *left censoring* occurs if a data point is below a certain value but it is unknown by how much. If the value of the censoring is a random variable, the random censoring time is usually assumed to be statistically independent of the failure time. More recently, however, the determination of failure times under *dependent* random censoring has been considered as well (Hsieh & Chen, 2020; Wang *et al.*, 2012).

Dependent censoring. In medical experiments on tumorigenicity, for example, the failure time of interest, T, is usually the time to tumor onset, which is commonly not observed. Instead, only (i) the death (or sacrifice) time of an animal,

16 FGM copula extensions with a larger dependency range exist but require additional parameters.

serving as the observation time X here, is observed and (ii) whether or not T exceeds the observation time X (at that time, one knows the absence or presence of the tumor). Thus, one can directly estimate the following two functions:

$$G(x) = P(X \leq x) \quad \text{and} \quad p_2(x) = P(X \leq x, T < X) \quad (0 \leq x \leq \infty)$$

by their empirical estimates. With $F(t)$ and $G(x)$ the distribution functions of T and X, respectively, one assumes a copula C

$$\mathrm{C}(F(t), G(x))$$

to specify the dependence between failure time and observation time. Importantly, it has been shown that, under weak assumptions and given the copula, the marginal distribution function F is *uniquely* determined by $G(x)$ and $p_2(x)$ (Wang *et al.*, 2012).

To show the formal equivalence with the dependent race model, we equate distribution $G(x)$ with $F_{go}(s)$ and $F(t)$ with $F_{stop}(t)$. Thus, $p_2(x) = P(X \leq x, T < X)$ corresponds to $P(T_{go} \leq s, T_{stop} + t_d < T_{go})$. Since the latter is not observable, we use the following equality:

$$\begin{aligned} &P(T_{go} \leq s) - P(T_{go} \leq s, T_{stop} + t_d < T_{go}) \\ &\quad = P(T_{go} \leq s, T_{go} < T_{stop} + t_d) \\ &\quad = P(T_{go} \leq s \mid T_{go} < T_{stop} + t_d)\, P(T_{go} < T_{stop} + t_d) \\ &\quad = F_{sr}(s \mid t_d)\, [1 - p_r(t_d)], \end{aligned}$$

showing a one-to-one correspondence between the observable quantities in dependent censoring and the stop-signal race model; note that we made use of the correspondence of $p_2(\infty)$ with $P(T_{stop} + t_d < T_{go}) \equiv p_r(t_d)$.

Consequently, the uniqueness result in dependent censoring implies that $F_{stop}(t)$ is uniquely determined in the general race model with a specified copula and that the distribution is amenable to nonparametric estimation methods developed in actuarial science (e.g., Titman, 2014 for a maximum likelihood method). This result is very general and applies to any dependent model, e.g., the FGM model defined in Equation (7.29). Unfortunately, however, a further well-known result from that theory implies that the *numerical value* of the dependence parameter, e.g., δ in the case of the FGM model, is not identifiable in general and thus cannot be estimated without specifying the marginals (Betensky, 2000; Titman, 2014). Nevertheless, a sensitivity analysis can be quite revealing about the impact of dependency (Wang *et al.*, 2012). In the FGM model, this involves taking a range of dependency parameter values, like $\delta = 0, \pm 0.1, \pm 0.2, \ldots, \pm 0.5$, and probing how much the predictions generated for the stop-signal distribution vary as a function of these values. An application of these results to empirical stop-signal data has not yet been undertaken, however.

We conclude this section with a model featuring extreme stochastic dependency not requiring any numerical parameters.

7.7.3 Perfect Negative Dependency Race Model

In order to resolve the paradox described above, of interacting circuits of mutually inhibitory neurons instantiating stop and go processes in spite of stochastically independent finishing times, we have suggested a race model with negative dependency between go and stop-signal processing times (Colonius & Diederich, 2018). It is based on the countermonotonicity copula expressing *perfect negative dependence* (PND) between T_{go} and T_{stop} and is completely parameter-free. The bivariate distribution is defined as

$$H^{-}(s,t) = \max\{F_{go}(s) + F_{stop}(t) - 1, 0\} \tag{7.30}$$

for all s,t ($s,t \geq 0$). It follows that the marginal distributions of $H^{-}(s,t)$ are the same as before, that is, $F_{go}(s)$ and $F_{stop}(t)$. Moreover, it can be shown that Equation (7.30) implies that

$$F_{stop}(T_{stop}) = 1 - F_{go}(T_{go}) \tag{7.31}$$

holds *almost surely*, that is, with probability 1. Thus, for any F_{go} percentile we immediately obtain the corresponding F_{stop} percentile as complementary probability and vice versa, which expresses perfect negative dependence between T_{go} and T_{stop}.[17]

Colonius and Diederich (2018) show that the PND race model is consistent with the empirical data patterns of the stop-signal task (Section 7.2) and that one can test the model, at least in principle, against stochastically independent race models. However, experimental studies of the model are not yet available. The PND model arguably constitutes the most direct implementation of the notion of "mutual inhibition" observed in neural data: any increase of inhibitory activity (speed-up of T_{stop}) elicits a corresponding decrease in "go" activity (slow-down of T_{go}) and vice versa.

7.8 Miscellaneous Aspects

7.8.1 Variants of the Stop-Signal Paradigm

Early on, some variants of the standard stop-signal task have been developed in an attempt to gain further insight into response inhibition mechanisms (Logan & Burkell, 1986). Data obtained from these studies are mainly discussed against the background of the independent or the interactive race model. Formal modeling approaches geared toward the specific task variants are rare, however. Here we sketch some results and point out future research goals.

Stop-change paradigm. In stop-change tasks, subjects are instructed to stop the originally planned go response and execute an alternative "change" response

17 The relation in Equation (7.31) is also interpretable as "T_{stop} *is (almost surely) a decreasing function of* T_{go}."

(or "go_2" task) when a signal occurs. A number of experimental and modeling studies suggest that subjects cannot stop and replace a response by simply activating an alternative response. A stop process must inhibit the first go response before the go_2 response can be executed. For some modeling efforts within the multitasking context, we refer to Verbruggen, Schneider, and Logan (2008).

Selective-stop paradigm. There are two variants of the selective stop task: in *stimulus-selective* stopping tasks, different signals can be presented and subjects must stop if one of them occurs (valid signal), but not if the others occur (invalid signals); in *motor-selective* stop tasks, subjects must stop some of their responses (e.g., finger press) but not others (e.g., foot press).

For the stimulus-selective task, there are three different types of trials: (i) only the go signal is presented; (ii) both the go signal and the stop signal are presented; and (iii) both the go signal and the ignore signal are presented. Mainly, two alternative strategies for stimulus-selective stopping have been discussed within the race model framework: "Stop then Discriminate" and "Discriminate then Stop" (Bissett & Logan, 2014). Given that stop and ignore signals are never presented within one and the same trial, it is not obvious that discriminating between stop and ignore signals can naturally be represented as a race. It has been suggested that in the "Discriminate then Stop" strategy discrimination interferes with go processing, violating the context independence assumption of the independent race model. For this paradigm, further theoretical and experimental work is clearly called for.

Anticipated response inhibition. In anticipated response inhibition (ARI) tasks, participants are required to make a planned response that coincides with a predictably timed event (typically a vertically filling bar) at a predefined stationary target (e.g., horizontal line on the bar). This predefined target requires participants to consistently prepare and initiate movement and is supposed to avoid the "strategic slowing" often observed in the ordinary stop-signal task even when subjects are asked to "respond as soon as possible." Experimental comparisons of ARI tasks with the ordinary stop-signal task suggest indeed that SSRT estimates show less bias with this version of response inhibition task (Leunissen *et al.*, 2017). However, a recent study finds violations of context independence due to the nature of the task and suggests a parametric model to take those into account (Matzke *et al.*, 2021).

7.8.2 Modeling Trigger Failures

All race models assume that go processing (T_{go}) is triggered by presenting the go signal, and stop processing (T_{stop}) by occurrence of the stop signal. However, sometimes no response is registered before the response deadline. These "go omissions" may be due, e.g., to distraction or a lack of attention. For the nonparametric independent race model, a recommendation by Verbruggen *et al.* (2019), based on extensive simulations, is to assign the maximum observed RT

in order to compensate for the lacking responses, when the integration method of estimating SSRT is used.

A more difficult problem arises if the stop signal fails to trigger the stopping process. Simulations have shown that nonparametric estimation methods will overestimate SSRT when trigger failures are present on stop trials (Band, van der Molen, & Logan, 2003). If the probability to respond in stop signal trials (the inhibition function) is larger than zero for small or zero SSDs, this suggests the presence of trigger failures. Unfortunately, there are typically only rather few observations available for very small SSDs, making estimates for this probability unreliable. As a pragmatic solution, Verbruggen *et al.* (2019) suggest researchers include extra stop signals that occur at the same time as the go stimulus but not include these trials in estimating SSRT.

At this time, there is no general solution available for nonparametric race models to estimate the probability of trigger failures in stop-signal trials. On the other hand, recent variants of parametric modeling methods provide an estimate of the probability of such trigger failures using a distribution-mixture approach (for details, see Matzke *et al.*, 2019).

7.8.3 Sequential (After Effects) Effects

In a large variety of action control tasks like the stop-signal paradigm, participants typically slow down after an error ("post-error slowing"). Several distinct behavioral and physiological explanations have been offered for this observation (Ullsperger, Danielmeier, & Jocham, 2014), but quantitative models are scarce (though see Dutilh *et al.*, 2012 for a diffusion model approach). One hypothesis attributes slowing to the "executive system": when it detects an error, it increases control by adjusting the parameters of the perceptual and response system to reduce the likelihood of committing future errors. Consistent with this, subjects often slow down after an unsuccessful stopping in the stop-signal task. However, slowing has been observed after successful stopping as well (Verbruggen & Logan, 2008). Bissett and Logan (2012) suggested that the presentation of the stop signal encourages subjects to shift priority from the go task to the stop task, producing longer response latencies after a signal trial and reducing the latency of the stop process. A formal approach has been undertaken by Soltanifar and colleagues (Soltanifar *et al.*, 2019). They estimate SSRT separately depending on whether the preceding trial has been a go or a stop trial and then develop a two-state mixture model for the SSRT distribution. They find clear effects of trial type, but further research along these lines is called for.[18] In an earlier development, Yu and colleagues suggested a comprehensive Bayesian inference-based, optimal-control theory for sequential effects (Ma & Yu, 2016) where a control system computes, at any given trial, the probability of a stop signal occurring in the next trial.

18 Unfortunately, in this and later papers, these authors always use parametric model versions only.

7.9 Concluding Remarks

In this chapter, we have aimed at characterizing the formal structure of quantitative models for the stop-signal task. Possible extensions and generalizations of currently available models have been discussed as well, e.g., the class of semi-parametric race models. Given the rapid increase of experimental studies in this area, presenting the empirical success (or failure) of the various models remains outside the reach of this chapter, however.

A recurring theme concerning model building is the issue of parametric versus nonparametric approaches. On the one hand, the independent, nonparametric race model of Logan and Cowan (1984) (Section 7.3.2), with its straightforward estimation methods for SSRT, has clearly dominated empirical studies up to now, notwithstanding numerous reports of violations of some of its assumptions. On the other hand, the availability of software packages for parameter estimation and model simulation is currently generating a broader usage of parametric race models in applied fields. Increased information about stop-signal processing time (beyond the mean), the possibility to more adequately deal with errors in choice paradigms that require discrimination between go signals, and the handling of stop-signal trigger failures have been listed among the benefits of the parametric approach (Matzke *et al.*, 2019). It should be mentioned for completeness, though, that it also faces some challenges. There is some arbitrariness involved in the choice of a specific family of distributions for go and stop-signal processing times (and, for Bayesian methods, in the choice of priors). For example, the commonly used ex-Gaussian distribution has some features that seem problematic: (i) it has an increasing hazard function, whereas most RT distributions exhibit an increasing and then decreasing (to some constant) hazard function (e.g., Luce, 1986, p. 439); and (ii) it predicts a nonzero probability of realizing negative values. The fact that ex-Gauss distributions often yield good empirical fits does not automatically mean that the ex-Gaussian parameters of the stop-signal distribution can be taken as valid description of the inhibitory process. Alternative distribution families have been considered, like the log-normal or the Wald distribution, but detailed studies have sometimes revealed broad parameter identifiability failures for these families (Matzke, Logan, & Heathcote, 2020).

It is difficult to predict what type of behavioral modeling will prevail in the future. In any case, it is obvious that the different variants of the paradigm, like selective stopping, will require going beyond the simple "race" scheme. Further insight from neurophysiology may suggest more complex mechanisms. A case in point is the two-stage pause-then-cancel (PTC) model by Schmidt and Berke (2017), based on subcortical rodent recordings. As described in Diesburg and Wessel (2021), the first stage is defined by a short-latency "Pause" process that actively delays the go process; it is followed by a slower "Cancel" process, which shuts off ongoing invigoration of the go response. This way, the PTC model tries to disentangle attentional orienting from motor inhibition. The model is clearly at odds with standard, independent race models and calls for an augmented

mathematical formalization with more sophisticated quantitative measures for the strength of inhibition.

7.10 Related Literature

While there are a number of early references to the stop-signal paradigm (e.g., Lappin & Eriksen, 1966), the first completely developed modeling approach is found in Logan and Cowan (1984). Over the years, a number of review articles have appeared, with different emphases (Band, van der Molen, & Logan, 2003; Logan, 1994; Logan *et al.*, 2014; Matzke *et al.*, 2018; Verbruggen & Logan, 2009; Verbruggen *et al.*, 2019). Platform-independent software to correctly execute the standard stop-signal task by F. Verbruggen is found on GitHub (https://github.com/fredvbrug/STOP-IT). For the anticipated response inhibition task, an open-source program (OSARI) is presented in He *et al.* (2021).

References

Band, G., van der Molen, M., & Logan, G. (2003). Horse-race model simulations of the stop-signal procedure. *Acta Psychologica*, *112*, 105–142.

Betensky, R. (2000). On nonidentifiability and noninformative censoring for current status data. *Biometrika*, *81*(1), 218–221.

Bissett, P., Jones, H., Poldrack, R., & Logan, G. (2021). Severe violations of independence in response inhibition tasks. *Science Advances*, *7*, eabf4355.

Bissett, P., & Logan, G. (2012). Post-stop-signal slowing: Strategies dominate reflexes and implicit learning. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 746–757.

Bissett, P., & Logan, G. (2014). Selective stopping? Maybe not. *Journal of Experimental Psychology: General*, *143*(1), 455–472.

Bogacz, R., Brown, E., Moehlis, P., Holmes, P., & Cohen, J. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Bompas, A., Campbell, A., & Sumner, P. (2020). Cognitive control and automatic interference in mind and brain: A unified model of saccadic inhibition and countermanding. *Psychological Review*, *127*(4), 524–561.

Bompas, A., & Sumner, P. (2009). Temporal dynamics of saccadic distraction. *Journal of Vision*, *9*(17), 1–14.

Bompas, A., & Sumner, P. (2011). Saccadic inhibition reveals the timing of automatic and voluntary signals in the human brain. *The Journal of Neuroscience*, *31*, 12501–12512.

Boucher, L., Palmeri, T., Logan, G., & Schall, J. (2007). Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review*, *114*(2), 376–397.

Brown, J., Hanes, D., Schall, J., & Stuphorn, V. (2008). Relation of frontal eye field activity to saccade initiation during a countermanding task. *Experimental Brain Research*, *190*, 135–151. (doi: 10.1007/s00221-008-1455-0)

Busemeyer, J., & Townsend, J. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459.

Carpenter, R. H., & Williams, M. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*, 59–62.

Colonius, H. (1990). A note on the stop-signal paradigm, or how to observe the unobservable. *Psychological Review*, *97*(2), 309–312.

Colonius, H. (2016). An invitation to coupling and copulas: With applications to multisensory modeling. *Journal of Mathematical Psychology*, *74*, 2–10. (http://dx.doi.org/10.1016/j.jmp.2016.02.004)

Colonius, H., & Diederich, A. (2001/2021). *Measuring the time to cancel a saccade.* (unpublished manuscript, available from https://uol.de/en/hans-colonius/my-publications)

Colonius, H., & Diederich, A. (2018). Paradox resolved: Stop signal race model with negative dependence. *Psychological Review*, *125*(6), 1051–1058.

Colonius, H., Özyurt, J., & Arndt, P. (2001). Countermanding saccades with auditory stop signals: Testing the race model. *Vision Research*, *41*(15), 1951–1968.

Diederich, A. (1995). Intersensory facilitation of reaction time: Evaluation of counter and diffusion coactivation models. *Journal of Mathematical Psychology*, *39*, 197–215.

Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, *41*, 260–274.

Diederich, A., & Mallahi-Karai, K. (2018). Stochastic methods for modeling decision-making. In W. H. Batchelder, H. Colonius, & E. Dzhafarov (Eds.), *New handbook of mathematical psychology*, (vol. 2, pp. 1–70). Cambridge: Cambridge University Press.

Diesburg, D., & Wessel, J. R. (2021). The pause-then-cancel model of human action-stopping: Theoretical considerations and empirical evidence. *Neuroscience & Biobehavioral Reviews*, *129*, 17–34.

Durante, F., & Sempi, C. (2016). *Principles of copula theory*. Boca Raton, FL: CRC Press.

Dutilh, G., Vandekerkhove, J., Forstmann, B., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics*, *74*, 454–465.

Forstmann, B., & Wagenmakers, E. (2015). *An introduction to model-based cognitive neuroscience*, 1st ed. Dordrecht: Springer.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/ hierarchical models*. Cambridge: Cambridge University Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741.

Hanes, D., & Carpenter, R. H. (1999). Countermanding saccades in humans. *Vision Research*, *39*, 2777–2791.

Hanes, D., Patterson, W., & Schall, J. (1998). Role of frontal eye field in countermanding saccades: Visual, movement and fixation activity. *Journal of Neurophysiology*, *79*, 817–834.

Hanes, D., & Schall, J. (1995). Countermanding saccades in macaque. *Visual Neuroscience*, *12*, 929–937.

He, J., Hirst, R., Puri, R., Coxon, J., Byblow, W., Hinder, M., … Puts, N. (2021). Osari, an open-source anticipated response inhibition task. *Behavior Research Methods*, https://doi.org/10.3758/s13428-021-01680-9.

Hsieh, J.-J., & Chen, Y.-Y. (2020). Survival function estimation of current status data with dependent censoring. *Statistics and Probability Letters*, *157*(108621).

Joe, H. (2015). *Dependence modeling with copulas* (Vol. 134). Boca Raton, FL: CRC Press.

Jones, M., & Dzhafarov, E. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*, 1–32.

Lappin, J. S., & Eriksen, C. (1966). Use of a delayed signal to stop a visual reaction-time response. *Journal of Experimental Psychology*, *72*(6), 805–811.

Leunissen, I., Zandbelt, B., Potocanac, Z., Swinnen, S., & Coxon, J. (2017). Reliable estimation of inhibitory efficiency: To anticipate, choose or simply react? *European Journal of Neuroscience*, *45*, 1512–1523.

Logan, G. (1994). On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In D. Dagenbach & T. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 189–239). San Diego, CA: Academic Press.

Logan, G. (2017). Taking control of cognition: An instance perspective on acts of control. *American Psychologist*, *72*(9), 875–884.

Logan, G., & Burkell, J. (1986). Dependence and independence in responding to double stimulation: A comparison of stop, change, and dual-task paradigms. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 549–563.

Logan, G., & Cowan, W. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, *91*(3), 295–327.

Logan, G., van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review*, *121*, 66–95.

Logan, G., Yamaguchi, M., Schall, J., & Palmeri, T. (2015). Inhibitory control in mind and brain 2.0: Blocked-input models of saccadic countermanding on the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review*, *122*, 115–147.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Ma, N., & Yu, A. (2016). Inseparability of go and stop in inhibitory control: Go stimulus discriminability affects stopping behavior. *Frontiers in Neuroscience*, *10:54*, doi: 10.3389/fnins.2016.00054.

Matzke, D. (2013). Releasing the BEESTS: Bayesian estimation of stop-signal reaction time distributions. *Frontiers in Quantitative Psychology and Measurement*, *4:918*, doi: 10.3389/fpsyg.2013.00918.

Matzke, D., Curley, S., Gong, C., & Heathcote, A. (2019). Inhibiting responses to difficult choices. *Journal of Experimental Psychology: General*, *148*(1), 124–142.

Matzke, D., Dolan, C., Logan, G., Brown, S., & Wagenmakers, E.-J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General*, *142*(4), 1047–1073.

Matzke, D., Logan, G., & Heathcote, A. (2020). A cautionary note on evidence-accumulation models of response inhibition in the stop-signal paradigm. *Computational Brain & Behavior*, *3*, 269–288.

Matzke, D., Strickland, L., Sripada, C., Weigard, A., Puri, R., He, J., … Heath, A. (2021). Stopping timed actions. *PsyArXiv*, https://doi.org/10.31234/osf.io/9h3v7.

Matzke, D., Verbruggen, F., & Logan, G. (2018). The stop-signal paradigm. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience: Methodology*, 4th ed. (Vol. 4, pp. 383–427). New York: Wiley.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817.

Middlebrooks, P., Zandbelt, B., Logan, G., Palmeri, T., & Schall, J. (2020). Countermanding perceptual decision-making. *iScience*, *23*(1), https://doi.org/10.1016/j.isci.2019.100777.

Nelsen, R. (2006). *An introduction to copulas*, 2nd ed. New York: Springer.

Özyurt, J., Colonius, H., & Arndt, P. (2003). Countermanding saccades: Evidence against independent processing of go and stop signals. *Perception & Psychophysics*, *65*(3), 420–428.

Paré, M., & Hanes, D. (2003). Controlled movement processing: Superior colliculus activity associated with countermanded saccades. *Journal of Neuroscience*, *23*, 6480–6489.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Reingold, E., & Stampe, D. (2002). Saccadic inhibition in voluntary and reflexive saccades. *Journal of Cognitive Neuroscience*, *14*, 371–388.

Schall, J. (2004). On building a bridge between brain and behavior. *Annual Review of Psychology*, *55*, 23–50.

Schall, J. (2019). Accumulators, neurons, and response time. *Trends in Neurosciences*, *42*(12), 848–860.

Schall, J., & Godlove, D. (2012). Current advances and pressing problems in studies of stopping. *Current Opinion in Neurobiology*, *22*, 1012–1021.

Schall, J., Palmeri, T., & Logan, G. (2017). Models of inhibitory control. *Philosophical Transactions of the Royal Society of London B*, *372*(20160193), http://dx.doi.org/10.1098/rstb.2016.0193.

Schall, J., & Paré, M. (2021). The unknown but knowable relationship between presaccadic accumulation of activity and saccade initiation. *Journal of Computational Neuroscience*, https://doi.org/10.1007/s10827-021-00784-7.

Schmidt, R., & Berke, J. (2017). A pause-then-cancel model of stopping: Evidence from basal ganglia neurophysiology. *Philosophical Transactions of the Royal Society of London B*, *372*, http://dx.doi.org/10.1098/rstb.2016.0202.

Smith, P. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408–463.

Smith, P., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116*, 283–317.

Soltanifar, M., Dupuis, A., Schachar, R. J., & Escobar, M. (2019). A frequentist mixture modeling of stop signal reaction times. *Biostatistics & Epidemiology*, *3*(1), 90–108.

Teller, D. (1984). Linking propositions. *Vision Research*, *24*(10), 1233–1246.

Titman, A. (2014). A pool-adjacent-violators type algorithm for non-parametric estimation of current status data with dependent censoring. *Lifetime Data Analysis*, *20*, 444–458.

Trappenberg, T., Dorris, M., Munoz, D., & Klein, R. (2001). A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *Journal of Cognitive Neuroscience*, *13*, 256–271.

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, *94*, 35–79.

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550–592.

van Zandt, T., Colonius, H., & Proctor, R. (2000). A comparison of two response-time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*, 208–256.

Verbruggen, F., Aron, A. R., Band, G., Beste, C., Bissett, P. G., Brockett, A. T., … Boehler, C. N. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *eLife*, *8*. doi: 10.7554/eLife.46323

Verbruggen, F., & Logan, G. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, *12*, 418–424. (http://doi.org/10.1016/j.tics.2008.07.005)

Verbruggen, F., & Logan, G. (2009). Models of response inhibition in the stop-signal and stop-change paradigms. *Neuroscience & Biobehavioral Reviews*, *33*, 647–661. (http://doi.org/10.1016/j.neubiorev.2008.08.014)

Verbruggen, F., Schneider, D., & Logan, G. (2008). How to stop and change a response: The role of goal activation in multi-tasking. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 1212–1228.

Walker, R., & Benson, V. (2013). Remote distractor effects and saccadic inhibition: Spatial and temporal modulation. *Journal of Vision*, *13*(9), 1–21.

Wang, C., Sun, J., Sun, L., Zhou, J., & Wang, D. (2012). Nonparametric estimation of current status data with dependent censoring. *Lifetime Data Analysis*, *18*, 434–445.

8 Approximate Bayesian Computation

Noah Thomas, Brandon M. Turner,
and Trisha Van Zandt

8.1 Introduction

As mathematical psychology has evolved from the development of mathematical representations of psychological experience and mathematical relationships between these representations and behavior (Coombs, Dawes, & Tversky, 1970; Krantz *et al.*, 1971) to computational models of behavior and brain function and the relationships between them (Farrell & Lewandowski, 2018; Liu *et al.*, 2020), the systems of analyses employed to analyze data and fit and test models have evolved to meet the increased demands of computation and complexity. A consequence of this complexity is reflected in the increased difficulty involved in using standard methods such as maximum likelihood to fit models to data. One powerful and straightforward way to circumvent these difficulties is Bayesian hierarchical modeling.

Bayesian modeling, as described in Rouder, Morey, and Pratte (2017), focuses attention on model parameters given a set of data. Its power, especially in the context of mathematical psychology, is in its ability to provide statements about

psychologically motivated parameters in the context of a theoretically interesting model. With a model stating that a vector of data $Y = y$ should follow some distribution $f(y \mid \theta)$, where the form of f and the model parameters θ are dictated by psychological theory, and with some initial assumptions about the prior distribution $\pi(\theta)$ of the parameters θ, we update our understanding of θ by

$$\pi(\theta \mid y) \propto f(y \mid \theta)\pi(\theta), \tag{8.1}$$

where the model's probability density function (pdf) or likelihood $f(y \mid \theta)$ is (roughly) the probability of the data set $Y = y$ arising under the parameter θ and the posterior distribution $\pi(\theta \mid y)$ is found by Bayes' rule.

A standard Bayesian analysis therefore requires that we are able to write down a function $f(y \mid \theta)$ that describes the random behavior of the data Y. This analysis often also assumes that for a random data set $Y = \{Y_1, Y_2, \ldots, Y_N\} = \{y_1, y_2, \ldots, y_N\}$, we write

$$L(\theta \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta), \tag{8.2}$$

or that the sample $Y = y$ is independent and identically distributed (i.i.d.): each Y_i is drawn from the same distribution (described by $f(y \mid \theta)$) and is independent from every other $Y_j, j \neq i$.

8.1.1 Increasing Sophistication of Models

Consider for a moment some of the hidden assumptions behind the i.i.d. assumption. First, we must assume that the data-generating process f is fixed in time. Neither the structure of the process (dictating the likelihood or distribution family f) nor its parameters (θ) change as a person gains experience or becomes bored with a task. Second, we assume that there are no serial position effects. Errors or slips on trial i have no effect on the execution of the process on trial $i + 1$. Third, we assume that all measurements come from the same process f; there are no contaminants from extraneous events, such as sneezing or hitting the wrong key. Fourth, we assume that everyone is the same: all individuals in the experiment do the task in exactly the same way, using the process f, perhaps also with exactly the same values of the parameter θ.

Many Bayesian models have also used a simplifying assumption of linearity, or that $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon$ for K predictor variables X_i and error ϵ that follows some distribution (usually Gaussian). This assumption brings with it a number of mathematical conveniences, such as the existence of conjugate priors and the direct relationships to methods of least squares.

The important contribution of the i.i.d. and linearity assumptions is that they usually result in a tractable form for Equation (8.2). The resulting models are useful, but oversimplified to the extent that the interpretation of the models' parameters is compromised (Haaf & Rouder, 2019). The Bayesian approach, however, is powerful enough that many of these assumptions are not necessary.

It is possible, using modern computing resources, to build and test models that are far more complex and realistic, as long as an explicit likelihood function $L(\theta \mid Y)$ can be formulated. Numerical methods based in Markov chain Monte Carlo sampling procedures exploit the theory of Markov processes, which assures that, under general conditions, simulating sequences of random numbers that have been accepted or rejected as potential samples from the desired posterior distribution will approach a sample from that posterior given enough time.

Such methods have been extensively addressed elsewhere (e.g., Gelman *et al.*, 2014; Rouder & Lu, 2005). The focus of this chapter is on situations in which the model of interest is so complex that the likelihood function $L(\theta \mid Y)$ is very difficult to determine or, indeed, doesn't exist at all. In this situation approximate Bayesian methods can be used to estimate the posterior distribution for parameters of models with no explicit likelihood $L(\theta \mid Y)$.

8.1.2 Statement of the Problem

There are a number of important, psychologically interesting models that do not have explicit likelihoods. For example, consider nonstationary models, in which the distribution of the observed data Y changes when the process that produces the data is shifted in time. This implies that $\theta(t)$, the model parameters, vary with time t. One example of a model like this is Usher and McClelland's leaky competing accumulator (LCA; Usher & McClelland, 2001). This model is based on a diffusion process framework in which the rate of evidence accumulation or drift changes as a function of the amount of evidence accumulated. As time passes, more evidence is accumulated, and so the drift rate changes. The LCA's drift rate is not constant within a trial and the resulting nonstationarity of the diffusion process means the data do not have an explicit likelihood.[1]

Models like the LCA are described by systems of nonlinear dynamic equations (see also the Orstein–Uhlenbeck process; e.g., Doob, 1942). Like neural network models, they must be simulated to determine their long-run behavior. Although such models have parameters θ that can be written down and given psychological interpretations (Busemeyer & Townsend, 1993), they do not have a likelihood $f(x \mid \theta)$ describing the random behavior of their outputs that can be analytically determined. Without a likelihood, how can we perform inference on model parameters, evaluate model fit, or contrast models to one another?

In what follows, we present approximate Bayesian analysis, the solution for models with no explicit likelihood. As a motivating example, we will consider a model for conflict tasks (described below), the activation-suppression race (ASR) model (Miller & Schwarz, 2021). Many models of conflict tasks, such as the diffusion model for conflict tasks (DMC; Ulrich *et al.*, 2015), are nonstationary and

1 Note that it is the dependence of the LCA model's parameters on time that makes the LCA nonstationary and eliminates the closed form of the model's likelihood function. The model produces i.i.d. observations.

do not have likelihoods. The ASR model does have a likelihood, so we can contrast standard MCMC methods with the approximate Bayesian methods to demonstrate the accuracy with which parameters can be estimated. We will describe the kinds of tasks that this model was designed to explain, and demonstrate, on simulated data, how the model can be fit to data and evaluated using approximate Bayesian analysis.

8.1.3 A Motivating Example: The Activation-Suppression Race Model of Conflict

A number of experimental paradigms in psychology make use of conflict tasks. The stimuli in such tasks are two-dimensional, carrying two sources of information. Each source contains information that is consistent with one of two possible responses. One example of a conflict task is the classic Stroop task (Stroop, 1935). Stroop stimuli are words of colors ("red," "green") printed in different inks. The participant's task is to identify the color of the ink. If the word "green" is printed in red ink, the correct response is "red," and the two sources of information in the stimulus are in conflict. If the word "red" is printed in red ink, the correct response is "red," and the two sources of information do not conflict. Participants respond faster and more accurately to stimuli without conflict than to stimuli with conflict.

The three most common conflict tasks are the Stroop task, the Eriksen flanker task (B. A. Eriksen & Eriksen, 1974), and the Simon task (Simon & Rudell, 1967). The flanker task uses a number of characters such as left- or right-pointing arrows arranged in a row or a column. Only the character in the center determines the response. If all the characters in the stimulus array point in the same direction, indicating the same response, then the two sources of information do not conflict; if the characters surrounding the critical central character indicate the opposite response, then the two sources of information conflict. One interesting feature of the flanker task is that the degree of conflict can be modulated by increasing or decreasing the separation between the characters, increasing or decreasing the number of flanking characters that are in conflict with the central character, and by increasing or decreasing the distance between the central character and the conflicting flankers (C. W. Eriksen & Hoffman, 1974; C. W. Eriksen & Schultz, 1977, 1978).

The Simon task asks that participants respond with a left keypress to a stimulus of one color (e.g., red) and a right keypress to a stimulus of a different color (e.g., green). In this task, conflict arises from the location of the stimulus relative to some central point (e.g., a central fixation cross) and the location of the desired response. If a red stimulus appears to the left of center, the two locations are not in conflict. If a green stimulus appears to the left of center, the two locations conflict. The degree of conflict can again be modulated by the distance of the stimulus from the center point. The more distal the stimulus, the faster participants' responses are to stimuli that conflict, and the slower they are to stimuli that don't conflict.

Miller and Schwarz (2021) proposed the ASR model to explain response times (RTs) in conflict tasks. The structure of the model is quite simple: two processes, A and B, are executed in parallel. The identification process B takes some amount of time T_B to determine the relevant information in a stimulus with or without conflict. The suppression process A takes some amount of time T_A to suppress information that is irrelevant for the correct response. When process B is finished, a third process C begins that selects and executes the response associated with the information that process B identified. Response selection and execution takes some amount of time T_C. The effect of conflict is observed on the response process C. If process A does not finish before process B, any irrelevant response information inhibits the response selection process C and prolongs the response by a duration λ_{inh}. Therefore

$$RT = \begin{cases} T_B + T_C & \text{if } T_A < T_B, \text{ otherwise} \\ T_B + T_C + \lambda_{\text{inh}} & \text{if } T_A > T_B \text{ and the stimulus conflicts.} \end{cases}$$

(If A finishes after B and the stimulus doesn't conflict there is no inhibition because there is no irrelevant response information in the stimulus.)

Letting T_A and T_B be exponentially distributed with rates α and β, respectively, and letting T_C be distributed as a Gaussian with mean μ_C and variance σ^2, the conditional likelihood for the RT is distributed as an ex-Gaussian variable with a mean that depends on the outcome of the race between A and B and the stimulus type. Noting that the probability p that B finishes before A is

$$p = \beta/(\alpha + \beta),$$

then, from Miller and Schwarz's (2021) expressions, the conditional RT distributions are

$$\begin{aligned} RT \mid \text{no conflict} \sim f_N(t \mid \theta) &= \text{exG}(\beta, \mu_C, \sigma) \text{ and} \\ RT \mid \text{conflict} \sim f_C(t \mid \theta) &= p\, \text{exG}(\alpha + \beta, \mu_C + \lambda_{\text{inh}}, \sigma) \\ &\quad + (1 - p)\left[(1 + \beta/\alpha)\text{exG}(\beta, \mu_C, \sigma)\right. \\ &\quad \left. - (\beta/\alpha)\text{exG}(\alpha + \beta, \mu_C, \sigma)\right], \end{aligned}$$

where $\text{exG}(\alpha, \mu_C, \sigma)$ is the ex-Gaussian density, the sum of an exponential (α) and a Gaussian (μ_C, σ) random variable, and $\theta = \{\alpha, \beta, \mu_C, \sigma, \lambda_{\text{inh}}\}$. Let an RT be denoted as the variable $T \in (0, \infty)$ and conflict as $C \in \{0, 1\}$, such that $C_i = 0$ if the stimulus presented on trial i does not conflict and 1 otherwise. Given n trials with and without conflict yielding a sample of RTs $T = \{T_1, T_2, \ldots, T_{2n}\}$, the ASR likelihood is

$$L(\theta \mid T) = \prod_{i=1}^{2n} (1 - C_i) f_N(T_i \mid \theta) + (C_i f_C(T_i \mid \theta)). \tag{8.3}$$

Because we can write down the ASR likelihood, we have no need to use likelihood-free methods. However, there are (common) situations where the likelihood is sufficiently complex that approximate Bayesian methods are easier to implement, either because it is difficult to code the likelihood or faster to simulate

the model. Models derived from stationary diffusion processes are examples of models with explicit likelihoods that are difficult to implement. For our purposes, we will perform both approximate and exact Bayesian computations for the ASR model to demonstrate that approximate Bayesian computations can result in accurate inference of model parameters.

8.2 Approximate Bayesian Computation

In 1984, Rubin discussed the need for applied statisticians to move beyond models incorporating simplifying assumptions, and to exploit modern computing resources to perform Bayesian analysis of less tractable models. While not strictly Bayesian, procedures that were "Bayesianly justifiable" were embraced by Rubin for their abilities to expand the range of models that could be applied to more complex problems. A simple simulation-based procedure for estimating posterior distributions, which he called "superpopulation frequency simulation," is now recognized as an example of approximate Bayesian computation (ABC).

Approximate Bayesian procedures were formalized by geneticists who were interested in the problem of determining how long ago the evolution of two species diverged from a common ancestor (Beaumont, Zhang, & Balding, 2002; Fu & Li, 1997; Pritchard *et al.*, 1999; Tavaré *et al.*, 1997; Weiss & von Haeseler, 1998). Representing the two species' genotypes as binary vectors, the locations and types of mutations along the vectors for each generation of the species must be explained by a model. There is no likelihood to describe the vector configurations over time, because the accumulation of possible mutations increases exponentially over time, and the length of the vectors and possible mutation sites increase as well.

Approximate methods that allowed for estimation of posterior distributions for the parameters of these models were reviewed by Beaumont, Zhang, and Balding (2002). Following publication of this review, the use of approximate Bayesian methods expanded from genetics to other disciplines, including psychology (Turner & Van Zandt, 2012). A comprehensive treatment of approximate Bayesian methods in cognitive modeling can be found in Palestro *et al.* (2018).

In what follows, we first outline the conceptual basis of approximate Bayesian analysis. Using the ASR as a motivating example, we then demonstrate three approaches for estimating the posterior distributions of the model's parameters and contrast those approaches with a standard Bayesian analysis. We outline the strengths and weaknesses of each approach.

8.2.1 Conceptual Basis

Rather than evaluating an explicit likelihood $f(x \mid \theta)$ for a data set T, approximate Bayesian methods depend on a comparison between an observed data set T and a data set T^* obtained by simulating the model of interest using a proposed set of parameters θ^*. We must define a distance metric $d(T, T^*)$ that quantifies the

discrepancies between T and T^*. While a true Bayesian posterior distribution is defined as $\pi(\theta \mid T)$ as in Equation (8.1), the approximate Bayesian posterior distribution is defined as $\pi(\theta \mid d(T, T^*) \leq \epsilon)$ for some small $\epsilon > 0$. If $d(T, T^*) \leq \epsilon$ for proposed parameters θ^*, we may retain θ^* as a sample from the desired posterior distribution of θ, otherwise we may discard it and select a new θ^*, simulate a new T^*, and repeat the process. Much depends on the magnitude of ϵ and how the distance $d(T, T^*)$ is defined. At its most precise definition, we could measure the Euclidean distance between the observed and simulated data vectors

$$\sqrt{\sum_{i=1}^{n}(T_i - T_i^*)^2},$$

and set $\epsilon = 0$. This would require each observation to be exactly reproduced in the simulation in the order in which it appeared, even if the T_is are an i.i.d. sample and order doesn't matter. The resulting sample of θ would come from the desired posterior distribution (an exact sample), but obtaining those samples would be computationally very expensive.

Recognizing that it is not necessarily the actual observations T^* that are required but some summary statistic(s) $S(T)$ of the sample that is required, the distance $d(T, T^*)$ can be redefined as $d(S(T), S(T^*))$. When deciding which summary statistic(s) to use, it would be theoretically optimal if the statistics used were *sufficient* for the model parameters θ. The *sufficiency principle* states that a sufficient statistic $S(T)$ provides as much information about a model parameter as the entire sample does. Examples of sufficient statistics include the sample mean

$$\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$$

and variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(T_i - \overline{T}\right)^2,$$

which are jointly sufficient for the mean μ_T and variance σ_T^2 of T.

Formally, $S(T)$ is sufficient for θ if the conditional distribution of the sample T given the value of the statistic $S(T)$,

$$\prod_{i=1}^{n} f(T_i \mid S(T)),$$

does not depend on θ.

Wilkinson (2013) demonstrated that exact samples from the desired posterior distribution can be obtained using approximate Bayesian methods so long as measurement error is additive and $d(T, T^*)$ is defined using sufficient statistics. However, if the model of interest does not have a likelihood $f(x \mid \theta)$, we won't be able to determine whether any given statistic $S(T)$ is sufficient for its parameters.

There is no guaranteed way around this problem, although certain techniques have been proposed (e.g., Fearnhead & Prangle, 2012). The most common approach is to use a larger number of statistics that accurately and adequately capture the shape of the likelihood distribution $f(x \mid \theta)$, such as the quantiles of T, keeping in mind that the more summary statistics we use, the more computation will be required to compute and evaluate the distance $d(T, T^*)$.

8.2.2 How it Works

Approximate Bayesian algorithms follow five general steps (Palestro *et al.*, 2018):

1. Generate a proposed value θ^* for θ. While the values of θ^* might be drawn from the prior distribution $\pi(\theta)$, this may not be the best choice, especially if the prior distribution is uninformative. The proposal distribution from which θ^* is drawn should not be too far away from the desired posterior distribution.
2. Simulate a data set T^* using θ^*. This step is computationally the most expensive. A data set must be simulated for every value of θ^* generated.
3. Compute the summary statistics of T^*. These could be sample moments, quantiles, or (in the case of the PDA algorithm discussed below) an estimate of the likelihood function.
4. Compare the summary statistics of T^* to those of T. This comparison may arise as computation of a distance d, or in some other, algorithm-specific way.
5. Weight θ^* according to how close T^* is to T. Because the simulation step is expensive, it is not always preferable to simply reject a θ^* when it produces T^* that is dissimilar to T. Instead, we might choose to assign a weight w to θ^* that indicates its fitness. How this weight is computed and translated into the posterior distribution will be determined by the algorithm selected.

Palestro *et al.* (2018) provide a comprehensive treatment of the choices that must be made in the implementations of a number of approximate Bayesian algorithms. Other treatments may be found in, e.g., Beaumont (2010); Cranmer, Brehmer, and Louppe (2020); Csilléry *et al.* (2010); Didelot *et al.* (2011). In this chapter we will present three of the most commonly used approximate Bayesian algorithms: rejection ABC, population Monte Carlo (PMC) sampling, and probability density approximation (PDA).

In what follows, we will use the notation identified in Table 8.1. All of the discussion will use the ASR model and RTs as measurements. The graphical ASR model is shown in Figure 8.1. All parameters were estimated on the log scale with Gaussian or improper uniform prior distributions. Table 8.2 gives the values of the Gaussian prior distributions.

To demonstrate the likelihood-free algorithms, we generated simulated data from the ASR. It is sufficient to discuss fits to a single simulated "participant" to demonstrate the key features of these algorithms. This participant's data were generated from a model with parameters equal to the values of the parameters in Table 8.2.

Table 8.1 *Notation used in the text.*

Symbol	Meaning
α	Rate parameter for the suppression process
β	Rate parameter for the identification process
λ_{inh}	The delay in response selection processing time induced by irrelevant information
μ_C	Mean of response selection processing time
σ^2	Variance of response selection processing time
θ	Vector-valued set of model parameters
θ^*	Set of proposed model parameters
θ_i	A sample of θ from its posterior distribution
w	A weight given to a proposed parameter value θ
T	Sample of n observed RTs
$T*$	Sample of simulated RTs
$S(T)$	A statistic computed from the sample T
n	The length of the data vector T
Model(θ)	The implicit conditional distribution of T observed through simulating a model with parameters θ, which takes the place of the likelihood function
$d(T, T^*)$	Distance between T and T^*
ϵ	A tolerance level for distance $d(T, T^*)$
$\pi(\theta)$	The prior distribution of θ
$\pi(\theta \mid T)$	The posterior distribution of θ
N	The number of iterations, corresponding to the number of samples of θ obtained from the estimated posterior distribution
$f(t \mid \theta)$	The density function of observation t for a model, also the likelihood for a single observation t
$L(\theta \mid T)$	The likelihood function for the data set T

Table 8.2 *Values of the parameters for the proper prior distributions of the ASR model.*

Parameter	Prior	Parameter	Prior
$\ln(\alpha)$	$\mathcal{N}(\ln 100, 0.75)$	$\ln(\beta)$	$\mathcal{N}(\ln 100, 0.75)$
$\ln(\mu_C)$	$\mathcal{N}(\ln 300, 0.3)$	$\ln(\sigma)$	$\mathcal{N}(\ln 60, 0.3)$
$\ln(\lambda_{\text{inh}})$	$\mathcal{N}(\ln 75, 1.0)$		

Each simulated RT was generated by first simulating durations for processes A and B by sampling values from exponential distributions with rate parameters α and β, respectively. Next, if the duration of process A was less than that of process B and/or the stimulus did not conflict, the duration for process C was simulated by sampling a value from a Gaussian distribution with mean μ_C and standard deviation σ. Finally, if the duration of process A was greater than that of process B and the stimulus conflicted, the duration of process C was simulated by

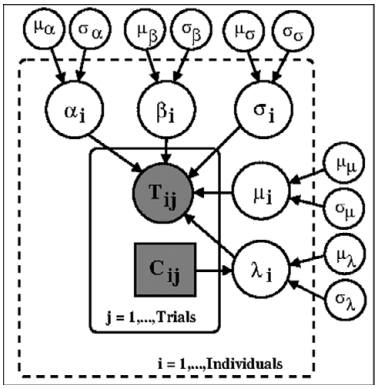


Figure 8.1 *Graphical model representation of the ASR model. Shaded shapes are observable variables, rectangles are integers, and circles are continuous quantities. Arrows represent dependencies, and plates represent repetition over trials (j) and individuals (i). The outermost plate (individuals) is dotted to indicate that this part of the graph is important for a hierarchical model, but we did not implement a hierarchical model for this exercise.*

sampling a value from a Gaussian distribution with mean $\mu_C + \lambda_{\text{inh}}$ and standard deviation σ.

8.2.3 Likelihood-Informed Markov Chain Monte Carlo

Before discussing the approximate Bayesian methods, we present the results from a standard Metropolis–Hastings MCMC algorithm (Martin, Quinn, & Park, 2011) to estimate the posterior distributions of the ASR parameter vector θ given a sample of 2,500 observations from a single simulated participant. The observations were generated for 1,250 stimuli that did have and 1,250 stimuli that did not have conflict. We computed the estimates using the MCMCmetrop1R() function from the R MCMCpack package.

Each parameter was modeled with an improper flat prior distribution on the log scale. That is

$$\pi\ (\ln \theta) \sim 1$$

for $\ln \theta \in \mathcal{R}$. The prior distribution is improper because it does not integrate to one, and the use of such a prior is equivalent to using no prior at all, placing all the importance on the data in the estimate of the posterior distribution. Because

$$\pi\ (\ln \theta) \propto L(\theta \mid X),$$

the posterior mode (the maximum *a posteriori* probability, or MAP estimate) is the maximum likelihood estimate of the parameter, and thus the usual caveats and concerns must be applied to the estimated posterior distributions (Bassett & Deride, 2019; Berger, 1985; Gelfand & Sahu, 1999; Hobert & Casella, 1996). In our case, the estimated posterior distributions are proper. We chose the improper flat prior where possible to put the different algorithms on footings that were as equal as possible. However, for certain algorithms a prior distribution must be

Data T, Model $T \sim \text{Model}(\theta)$;
prior distribution $\pi(\theta)$, tolerance ϵ;
Number of iterations N;
Initialize distance $d \leftarrow$ a big number;
for $i \leq N$ **do**
  Sample $\theta^* \sim \pi(\theta)$;
  Generate T^* from Model(θ^*);
  Compute $d \leftarrow d(T, T^*)$;
  if $d < \epsilon$ **then**
    Perform a Metropolis–Hastings accept/reject step;
  end
  Set $\theta_i \leftarrow \theta^*$ if accepted;
end

Algorithm 1 The rejection-based ABC algorithm to sample values of the parameter vector θ from its estimated posterior $\pi(\theta \mid T)$.

imposed and in these cases we employed the same diffuse prior distributions. In practice, improper (and indeed, non-informative) priors should be avoided unless considerable care is taken (Jaynes, 2003, Chapter 15), and researchers should perform sensitivity analyses (fitting a model with different priors) to evaluate the effects of the prior choice.

Figure 8.2 shows the estimated posterior distributions for the ASR parameters on the log scale, together with the improper priors shown as horizontal lines at 1.0. The true values of the parameters used to simulate the data are shown as vertical lines, and the 95% equal-tail credible sets are shown as the bars under the x-axis. The estimated exponential, delay, and Gaussian parameter posterior distributions are all centered close to the true value of the parameters that generated the data, and the true values are contained within the 95% credible sets.

We will use these estimated posterior distributions to evaluate the posterior distributions estimated using ABC methods.

8.3 Three ABC Algorithms

In this section, we present the three ABC algorithms that we selected to demonstrate the approximate Bayes concept. We begin with simple rejection, then a more complex population Monte Carlo procedure, and finally the probability density approximation algorithm. The results from each exercise demonstrate that the probability density approximation method yields the best approximation to the estimated posterior distributions recovered using the explicit likelihood function.

8.3.1 Rejection ABC

The easiest ABC algorithm to understand is the rejection algorithm, originally proposed by Pritchard *et al.* (1999) and shown in Algorithm 1. It is very easy to code, but can be very inefficient. It is also easy to extend to hierarchical models.

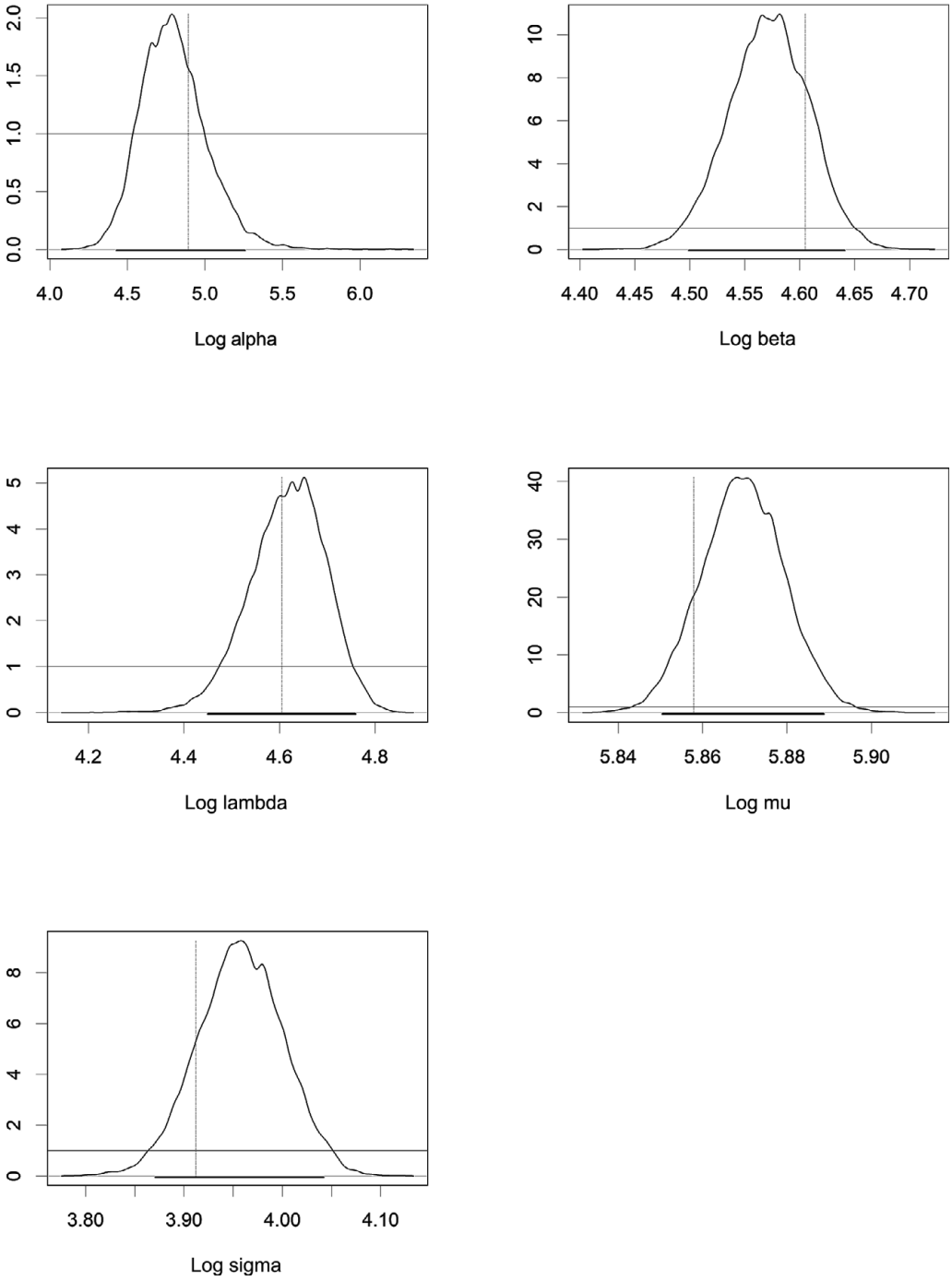


Figure 8.2 *Estimated posterior distributions (solid lines) and priors (horizontal gray lines) for the logged ASR parameters (α, β, λ_{inh}, μ_C, and σ, filled left to right, top to bottom) obtained using likelihood-informed Markov chain Monte Carlo. Values of the parameters used for simulation are shown as vertical dotted lines. Heavy horizontal lines at 0 show the 95% credible sets.*

To begin, we must first select the distance function $d(T, T^*)$ and a tolerance value ϵ. As we discussed earlier, these decisions are not necessarily easy to make, and most of the effort that goes into implementing the algorithm lies in figuring out what distance $d(T, T^*)$ and tolerance ϵ are most appropriate for a given Model(θ). Because $d(T, T^*)$ should ideally be a function of statistics sufficient for θ, and given that the entire sample is itself a sufficient statistic, one approach is to use distances that employ representations of the distribution of the sample such as quantiles. In this chapter, we use the χ-squared statistic by computing the quantiles Q of the observed (simulated) data (from 5% to 95%) and counting the number of observations in the proposal data T^* that fall in the bins defined by the quantiles Q. So, letting $f_i^*(T^*, Q)$ be the frequency of observations in T^* falling in the interval $[Q_{i-1}, Q_i)$:

$$\begin{aligned} d(T, T^*) &= \sum_{i=1}^{10} \frac{(f_i(T^*, Q) - O_i)^2}{O_i} \\ &= \frac{(f_1(T^*, Q) - 0.05n)^2}{0.05n} + \sum_{i=2}^{9} \frac{(f_i(T^*, Q) - 0.1n)^2}{0.1n} \\ &\quad + \frac{(f_{10}(T^*, Q) - 0.05n)^2}{0.05n}, \end{aligned}$$

where $Q_0 = 0$ and $Q_{10} = \infty$.

Selecting the tolerance value ϵ will require some exploration of the range of the model's predictions. Frequently we will select ϵ such that the proportion of rejected/accepted proposals is neither too high nor too low.[2] It may be difficult to find initial values of θ^* that produce samples T^* to satisfy the tolerance criterion ϵ. For this reason it may be more efficient to start with a large value for ϵ and then gradually reduce it to some minimum value.

Because the distance $d(T, T^*)$ is a function of θ^*, it will also be efficient to select values of θ^* that are close to those that satisfy stricter and stricter criteria (as tolerance ϵ is reduced). We can accomplish this by incorporating an MCMC sampling step (Robert & Casella, 2014). For example, we might apply a random walk that moves the samples of θ through the parameter space according to a Markov chain. So, for sample j:

$$\theta^* \sim \mathcal{N}(\theta_{j-1}, \sigma^*),$$

where $\mathcal{N}$ is the Gaussian distribution, θ_{j-1} is the most recently sampled value for θ, and σ^* is a tuning parameter that balances the need to search the parameter

2 Another alternative is to determine what the range of values of $d(T, T^*)$ will be when the model is true and select ϵ on that basis. This may be accomplished by simulating a data set T from the model using a set of parameters θ, and then simulating a large number of other data sets (call them $T^{(i)}$, $i = 1, \ldots, n$) using θ. For each simulated data set we can compute $d(T, T^{(i)})$ to obtain a distribution for $d(T, T^*)$. The tolerance value ϵ should not be smaller than the smallest of those values observed in the distribution of $d(T, T^*)$. This can be difficult if the distribution of $d(T, T^*)$ changes greatly with different values of θ^*.

space with the need to stay close to values of θ that satisfy the tolerance criterion ϵ. A good rule of thumb is to set σ^* equal to a value that produces an acceptance rate that is neither too high nor too low: around 23% (Roberts, Gelman, & Gilks, 1997).

We implemented the rejection ABC algorithm to estimate the ASR model's parameters. The starting values for each parameter were initialized at the means of their independent diffuse prior distributions:

$$\begin{aligned} \ln(\alpha) = \ln(\beta) &= 100, \\ \ln(\mu_C) = 300,\ \ln(\sigma) &= 60, \text{and} \\ \ln(\lambda_{\text{inh}}) &= 75. \end{aligned}$$

Following this initialization, on iteration i, new proposals were drawn from an independent Gaussian proposal distribution $q(\theta^* \mid \theta_i)$ with variance parameter $\sigma = 1.3$. The function q is often called a transition kernel. A kernel is a symmetric, non-negative function that integrates to 1, and the transition takes values of θ_i to values of θ_{i+1}.

We set the tolerance ϵ to 65. If a proposal θ^* produced a value of $d(T, T^*)$ that was less than ϵ, we evaluated it with a standard Metropolis–Hastings step, otherwise we rejected it. First we calculated the distance $d(T, T^*)$, and then

$$a = \begin{cases} \min\left(1, \dfrac{\pi(\theta^*)q(\theta_i \mid \theta^*)}{\pi(\theta_i)q(\theta^* \mid \theta_i)}\right) & \text{if } d(T, T^*) \leq \epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{8.4}$$

The value of a gives the (Bernoulli) probability of accepting θ^* given that $d(T, T^*)$ was less than ϵ. If θ^* was accepted, then $\theta_{i+1} = \theta^*$. We repeated the procedure and generated a chain of 1,000,000 samples of θ from its estimated posterior distribution.

As before, Figure 8.3 shows the estimated posterior distributions for the ASR parameters on the log scale together with their prior distributions. The true values for the parameters used to simulate the data are shown as vertical lines, and the 95% equal-tailed credible sets are shown as bars under the x-axis.

The estimated exponential, delay, and Gaussian parameter posteriors are all centered close to the true value of the parameters that generated the data, and the true values are contained within the 95% credible sets. However, the estimated posteriors are much broader that those resulting from the standard MCMC method. There could be many reasons for this discrepancy, including the value of the tolerance ϵ.

Theoretically, if $\epsilon = 0$, the approximations of the posterior distributions would be exact if the distance $d(T, T^*)$ is defined with sufficient statistics. If ϵ is too large, the estimated posterior variance will be large (as observed), because many more proposed values for θ that are distant from the MAP estimate will satisfy the tolerance criterion. If the quantile statistics defining $d(T, T^*)$ are not sufficient, the

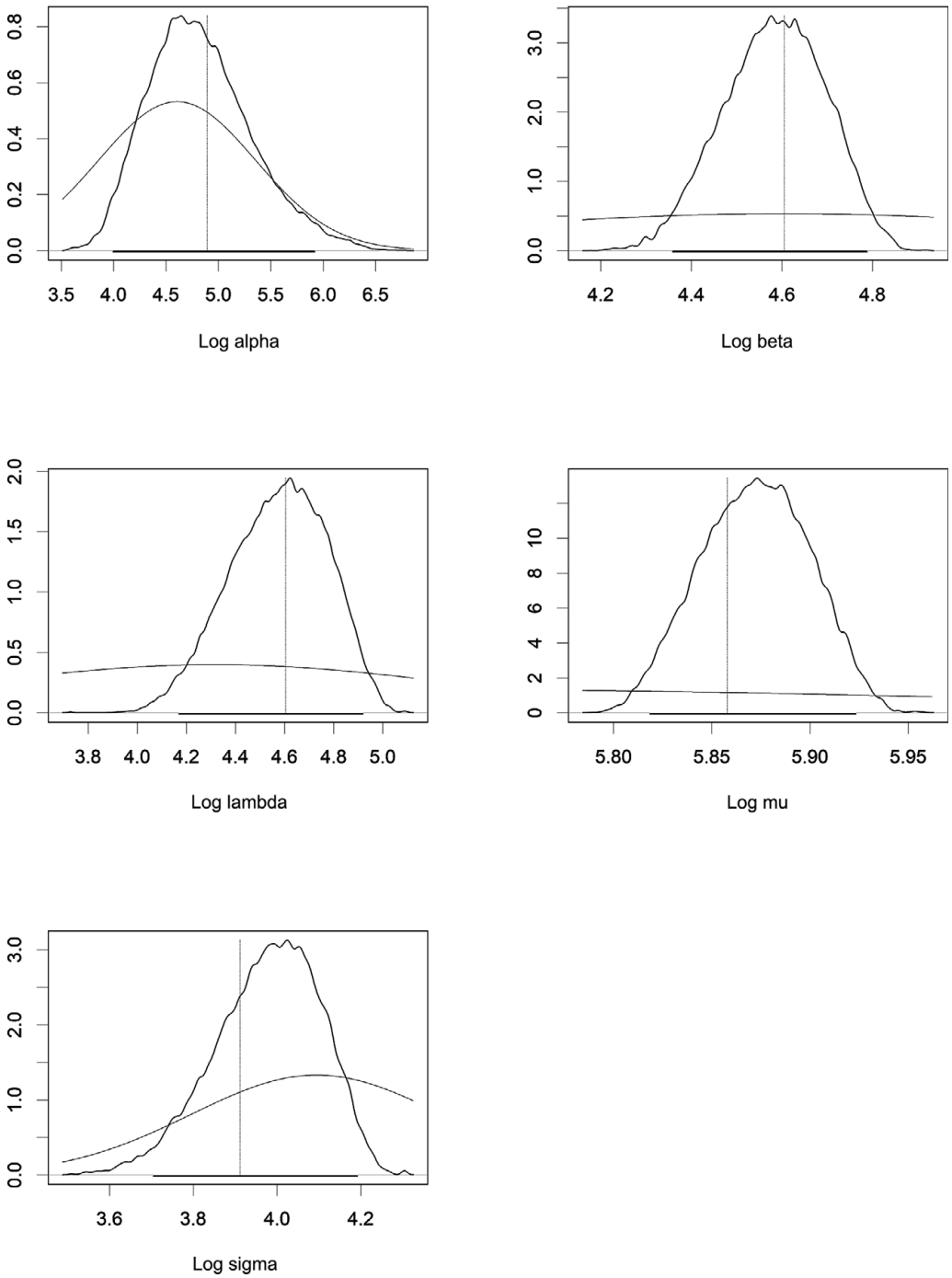


Figure 8.3 *Estimated posterior (solid lines) and prior (gray lines) distributions for the logged ASR parameters (α, β, λ_{inh}, μ_C, and σ, filled left to right, top to bottom) obtained using rejection Markov chain Monte Carlo. Values of the parameters used for simulation are shown as vertical dotted lines. Heavy horizontal lines at 0 show the 95% credible sets.*

estimates will also be inaccurate but it is difficult to predict how that inaccuracy would manifest. For the present exercise, it was impractical to set ϵ much smaller than 65; the samples that produced values of $d(T, T^*)$ less than 65 were rare, resulting in extraordinarily long estimation times.

Although the rejection algorithm is easy to implement and easy to understand, it has at least three significant drawbacks. First, it can be computationally very expensive if the prior distributions are far from the posterior, or if the tolerance criterion ϵ is poorly chosen. Second, the accuracy of the posterior estimate depends on the minimum value of ϵ selected and the sufficiency of the statistics defining $d(T, T^*)$. It may not be possible to achieve distances less than very small ϵ values. There is therefore a significant computational load versus accuracy tradeoff that may be difficult to resolve in an optimal way. The larger the minimum value of ϵ, the greater the variance of the estimated posterior distributions will be. Third, the MCMC sampler used to generate values of θ^* must be chosen with care. An improperly tuned sampler with a poorly chosen transition kernel will also contribute to inaccurate estimates.

Finally, it is far more difficult to fit hierarchical models using rejection ABC. The problem again rests with the tolerance criteria: values of ϵ must be set for each individual, given that different samples may result in larger distances. It becomes highly likely that the parameter chains for individuals become "stuck" in local minima, requiring adjustments to the ϵ values that are difficult to optimize in an automatic way (Turner & Van Zandt, 2014). Palestro *et al.* (2018) recommend block sampling of hyperparameter posterior distributions using (for example) Gibbs sampling, followed by sampling of individual parameters.

Nonetheless, if all that is required is a MAP estimate of a parameter, the rejection ABC method could be used to quickly obtain one that is reasonably close to the true MAP estimate (though with potentially high expected prediction error).

8.3.2 Population Monte Carlo

The PMC algorithm (see Algorithm 2) is based on a technique called particle filtering (Gordon, Salmond, & Smith, 1993). Instead of starting with a single initial value for θ, particle filtering algorithms generate a large set of values (a population). On each iteration of the algorithm, each value, or particle, is perturbed, evaluated for fitness, and accepted or rejected (filtered). The perturbation takes the form of a transition kernel, as described for the rejection algorithm. Fitness of each particle is evaluated using a distance function $d(T, T^*)$, and, based on that fitness, an importance weight is computed that determines the particle's probability of being accepted for the next iteration. Particles that are less fit have importance weights that, over time, become small, resulting in the particle being dropped out of the population. At the end of the algorithm, the population of particles that remain are a sample from an estimate of the desired posterior distribution (Cappé *et al.*, 2004).

Implementing the PMC algorithm requires a decision about how the importance weights w are to be computed. On the first iteration, all weights are equal to

Data T, Model $T^* \sim \text{Model}(\theta)$;
prior distribution $\pi(\theta)$, tolerance $\epsilon_{1:N}$;
Number of iterations N, number of particles M;
Iteration $j \leftarrow 1$;
Initialize distance $d \leftarrow$ a big number;
for $1 \leq i \leq M$ **do**
  while $d > \epsilon_1$ **do**
    Sample θ^* from the prior $\pi(\theta)$;
    Generate T^* from Model(θ^*);
    Compute $d \leftarrow d(T, T^*)$;
  end
  Set $\theta_{i,1} \leftarrow \theta^*$;
  Set $w_{i,1} \leftarrow 1/M$;
end
for $2 \leq j \leq N$ **do**
  for $1 \leq i \leq M$ **do**
    while $d > \epsilon_j$ **do**
      Sample θ^* from the pool θ with weights w;
      Perturb θ^* by sampling $\theta^{**} \sim \mathcal{N}(\theta^*, \sigma^*)$;
      Generate T^* from Model(θ^{**});
      Compute $d \leftarrow d(T, T^*)$;
    end
    $\theta_{i,j} \leftarrow \theta^{**}$;
    Calculate $w_{i,j}$;
  end
end

Algorithm 2 *Population Monte Carlo sampling algorithm.*

$1/M$ and so all particles have an equal chance of being selected. On all following iterations, denote the weight given to particle i on iteration t, $\theta_{i,t}$, as $w_{i,t}$, where

$$w_{i,t} = \pi(\theta_{i,t}) / \sum_{j=1}^{M} w_{j,t-1} q(\theta_{j,t-1} \mid \theta_{i,t}, \sigma_{t-1})$$

and

$$\sigma_{t-1}^2 = 2 \sum_{i=1}^{M} (\theta_{i,t} - \overline{\theta_{t-1}})^2 / M,$$

with

$$\overline{\theta_t} = \sum_{i=1}^{M} \theta_{i,t} / M.$$

So σ^2_{t-1} is twice the sample variance of the population θ_{t-1}. This weighting scheme results in a transition kernel that minimizes the Kullback–Leibler distance between the posterior distribution and the proposal distribution, optimizing the acceptance probability (Douc *et al.*, 2007).

We chose the same prior distributions as in the rejection algorithm described above. We selected a Gaussian distribution with variance σ^2_{t-1} for the transition kernel $q(\theta^* \mid \theta)$. An initial population of $M = 100$ particles was selected from the prior distribution that produced data T^* that gave an initial distance less than $\epsilon_1 = 100$. Using the χ-squared distance metric, we iterated through the population 1,000 times while linearly decreasing the distance metric to $\epsilon_{1,000} = 25$.

Figure 8.4 shows the estimated posterior distributions for the ASR parameters (α and β, and the delay parameter λ_{inh}, top row, and μ_C and σ, bottom row). The true values for the parameters used to simulate the data are shown as vertical lines, and the 95% equal-tailed credible sets are shown as bars under the x-axis.

In contrast to the rejection algorithm, the PMC algorithm is more difficult to implement. In addition, it has a number of disadvantages. First, there is no guarantee that populations will not get trapped in a local minima, as our example makes quite clear. The multimodal nature of the estimates of the posterior distributions is indicative of the populations' tendency to get trapped in different areas of the parameter space. Second, there is much trial and error in selecting both an appropriate population size and the function that determines how the tolerance ϵ decreases. Third, as in the rejection methods, decreasing tolerance ϵ results in an increase in computation time. Fourth, like rejection ABC, extending the algorithm to a hierarchical structure will increase computation time.

A practical advantage of PMC over rejection methods is that at any time in the filtering process the currently accepted population is an approximation to the desired posterior distribution. This approximation will improve with further iterations, but can be used to monitor changes in the estimate of the posterior distribution as the algorithm iterates.

8.3.3 Probability Density Approximation

The probability density approximation (PDA) procedure (Turner & Sederberg, 2014) is unique among ABC algorithms in that it does not depend on computing a distance between an observed and a simulated data set. Instead, it uses a nonparametric estimate of the likelihood in the form of the empirical density estimate computed from the simulated data. This density estimate might take a number of forms depending on the nature of the data, such as a histogram or a kernel estimate (Silverman, 1986). For a sample T^* of size n, we write

$$\hat{f}(t \mid T^*) = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{t - T_i^*}{h}\right),$$

for a function K (the kernel) and a tuning parameter h. The function K weights the values of T^* according to their distance from t. The tuning parameter h determines

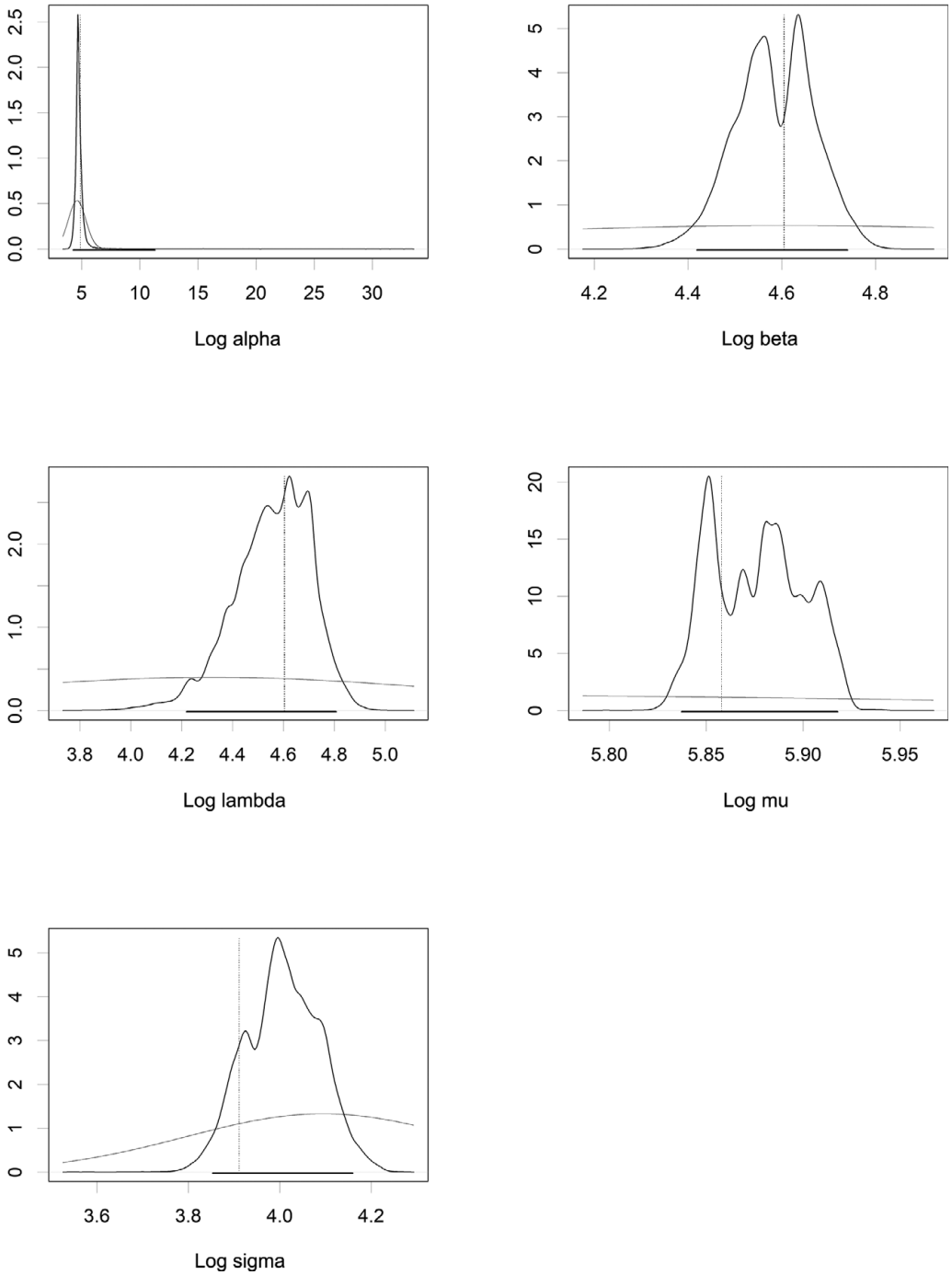


Figure 8.4 *Estimated posteriors (solid lines) and priors (gray lines) for the logged ASR parameters (α, β, λ_{inh}, μ_C, and σ, filled left to right, top to bottom) obtained using the population Monte Carlo algorithm. Values of the parameters used for simulation are shown as vertical dotted lines. Heavy horizontal lines at 0 show the 95% credible sets.*

Data T, kernel function K; prior distribution $\pi(\theta)$;
Number of iterations N; Initialize θ_1;
for $2 \leq i \leq N$ **do**
  $\theta^* \sim \pi(\theta)$;
  $T^* \sim$ Model(θ^*);
  Compute h^*;
  Compute $\hat{f}(t \mid T^*)$;
  Perform a Metropolis–Hastings accept/reject step;
  If accepted: Set $\theta_i \leftarrow \theta^*$
  Else: Set $\theta_i \leftarrow \theta_{i-1}$;
end

Algorithm 3 *Probability density approximation.*

how far a value of T^* can be from t and still influence the estimate of the density at the point t. There are a number of "plug-in" equations for h, and the most popular is Silverman's rule of thumb:

$$h = 0.9 \min\left(s_{T^*}, \frac{IQR}{1.34}\right) n^{-1/5},$$

where s_{T^*} is the (sample) standard deviation of T^* and IQR is the interquartile range of T^*.

The PDA algorithm is outlined in Algorithm 3. For each proposal θ^*, we simulated a data set T^* and constructed an empirical density estimate $\hat{f}(t \mid T^*)$. Given the prior distribution $\pi(\theta)$, we then performed a Metropolis–Hastings step by computing

$$a = \min\left(1, \frac{\pi(\theta^*)\hat{f}(t \mid t^*)q(\theta_{i-1} \mid \theta^*)}{\pi(\theta_{i-1})\hat{f}(t \mid T^*_{i-1})q(\theta^* \mid \theta_{i-1})}\right).$$

Noting that $\hat{f}(t \mid T^*)$ is an estimate of the likelihood $f(t \mid \theta^*)$, for a symmetric transition kernel $q(\theta_{i-1} \mid \theta^*)$ the numerator of a is an estimate of the marginal probability of the data T under θ^*, and the denominator is an estimate of the marginal probability of the data under θ_{i-1}. If the marginal probability of the data is greater under θ^* than under θ_i, then the new value of θ^* is accepted. If the marginal probability of the data is less, then the new value of θ^* is accepted with a probability that decreases as a function of how much less the marginal probability of T is under θ^*.

As in the likelihood-based MCMC estimation, we chose to use flat, improper priors to implement the PDA algorithm. The chain for θ was initialized at the values of the means in Table 8.2 and iterated 50,000 times.

Figure 8.5 shows the estimated posterior distributions for the logged exponential parameters (α and β) and the logged delay parameter (λ_{inh}, top row), and the estimated posterior distributions for the logged Gaussian parameters μ_C and σ (bottom row), together with the model's prior distributions. The true values for the parameters used to simulate the data are shown as vertical lines, and the 95%

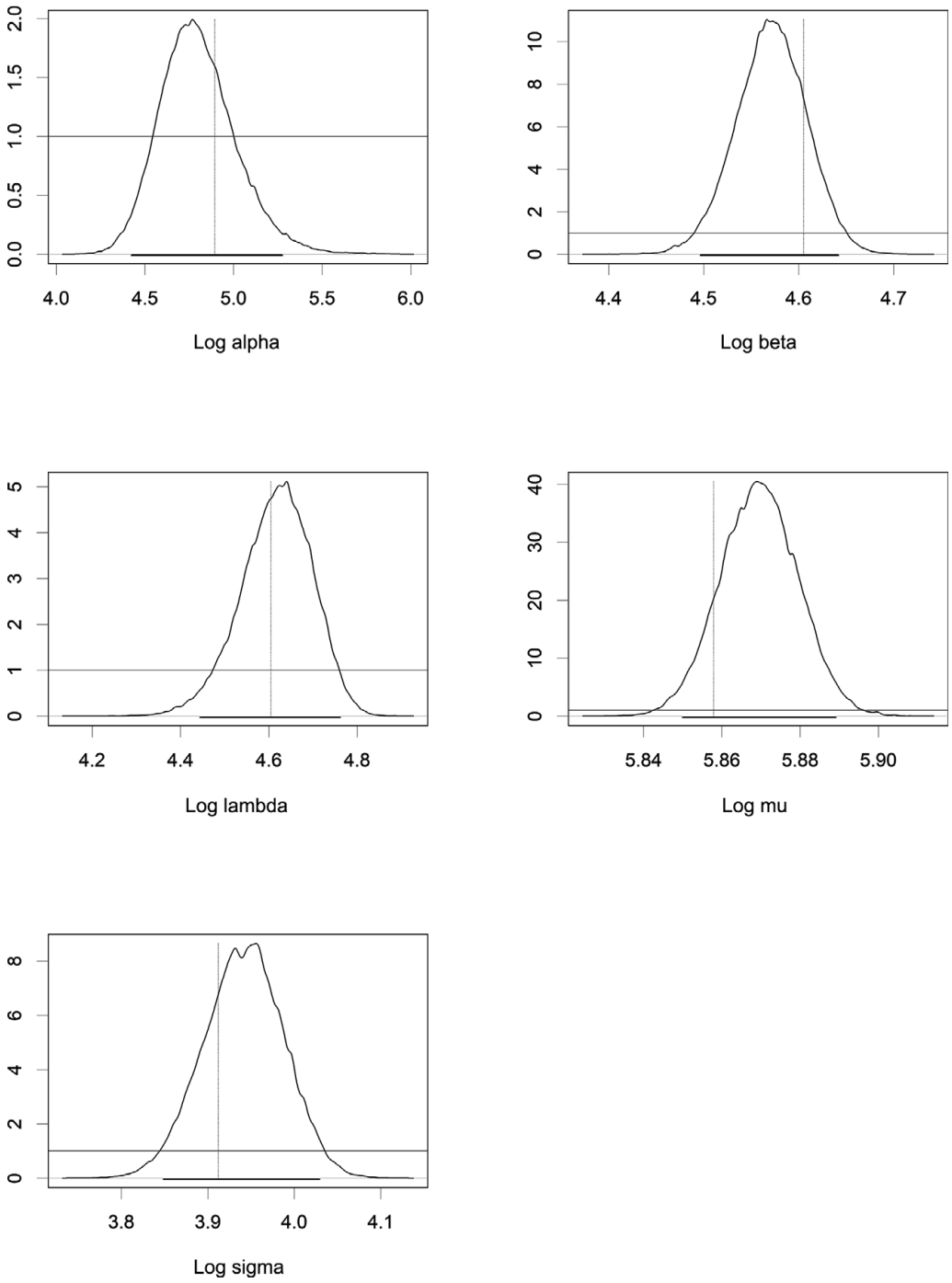


Figure 8.5 *Estimated posterior (solid lines) and prior (gray lines) distributions for the logged ASR parameters (α, β, λ_{inh}, μ_C, and σ, filled left to right, top to bottom) obtained using probability density approximation. Values of the parameters used for simulation are shown as vertical dotted lines. Heavy horizontal lines at 0 show the 95% credible sets.*

equal-tailed credible sets are shown as bars under the x-axis. Each credible set contains the true parameter value, and the posteriors are centered close to these true values.

8.3.4 Summary Results

Figure 8.6 is a violin plot of the estimated posterior distributions for the logged exponential parameters α and β, and the logged delay parameter λ_{inh} (top row) and the logged Gaussian parameters μ_C and σ (bottom row). Each violin corresponds to the posterior distribution estimated under each algorithm, identified on the x-axis.

To more precisely evaluate the ability of the PDA method to recover the posterior distributions estimated using standard MCMC methods, Figure 8.7 shows the quantile–quantile plots contrasting the posterior distributions of the ASR model obtained using MCMC and an explicit likelihood to those obtained using the PDA method. The top row of the figure contrasts the quantiles of the logged exponential parameters α and β and the logged delay parameter, and the bottom row contrasts the posterior quantiles of the logged Gaussian parameters μ_C and σ. While relatively obvious from Figure 8.6, these plots show that the results from the likelihood-informed MCMC approach are best approximated by the PDA algorithm, compared to the other algorithms we explored. In addition, the posterior mean of $\log \sigma$ is closer to the true parameter value using the PDA algorithm than the likelihood-based MCMC algorithm. This does not say anything about which estimates are more accurate, however, as the location of the posterior mean depends both on the data and the prior distributions. Thus the difference between the posterior mean and the value of the parameter that generated the data is not a complete picture of how accurate the posterior estimate is.

8.4 Conclusions

In this chapter we have discussed the approximate Bayesian method, and demonstrated three different algorithms, each a representative of a major ABC approach (rejection, particle filtering, and a probability density estimation technique). Each approach has strengths and weaknesses, ranging from ease of coding, computation time, and estimation accuracy.

Rejection approaches are easy to code. The accuracy of the final estimates of the posterior distributions obtained with rejection methods depend on the distance measure and tolerance criterion ϵ chosen. High accuracy with these methods usually requires a large sacrifice in computation time. As ϵ goes to zero, it is very difficult to find parameter values that generate data that can satisfy that criterion. If the criterion is too high, computation time can be very fast, but the estimated posterior distributions will be over-dispersed.

Particle filtering methods are more difficult to code, but can result in more accurate estimates for some applications (Cappé, Godsill, & Moulines, 2007).

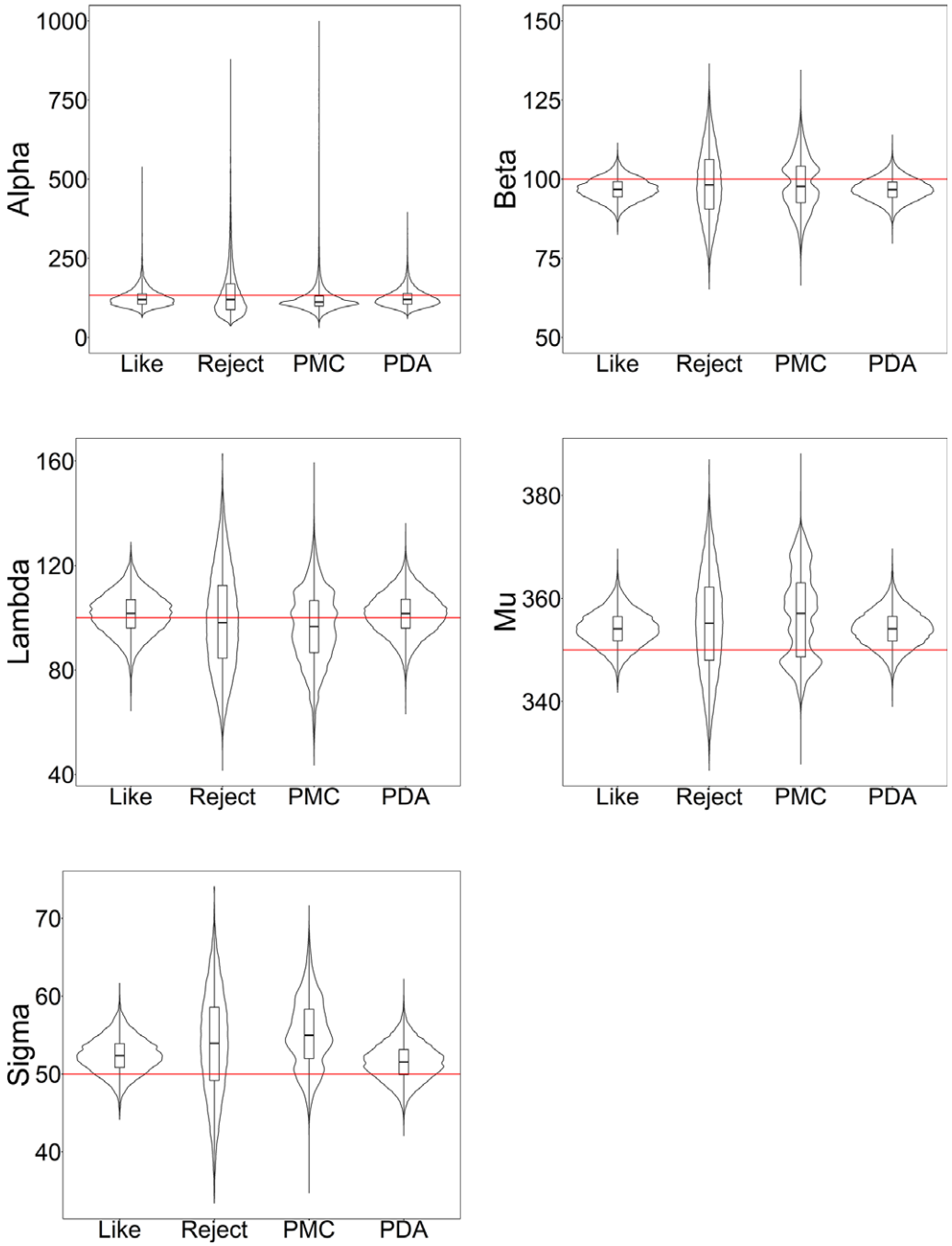


Figure 8.6 *Estimated posterior distributions for the logged ASR parameters (α, β, λ_{inh}, μ_C, and σ, filled left to right, top to bottom) obtained for each algorithm. The algorithm is shown on the x-axis.*

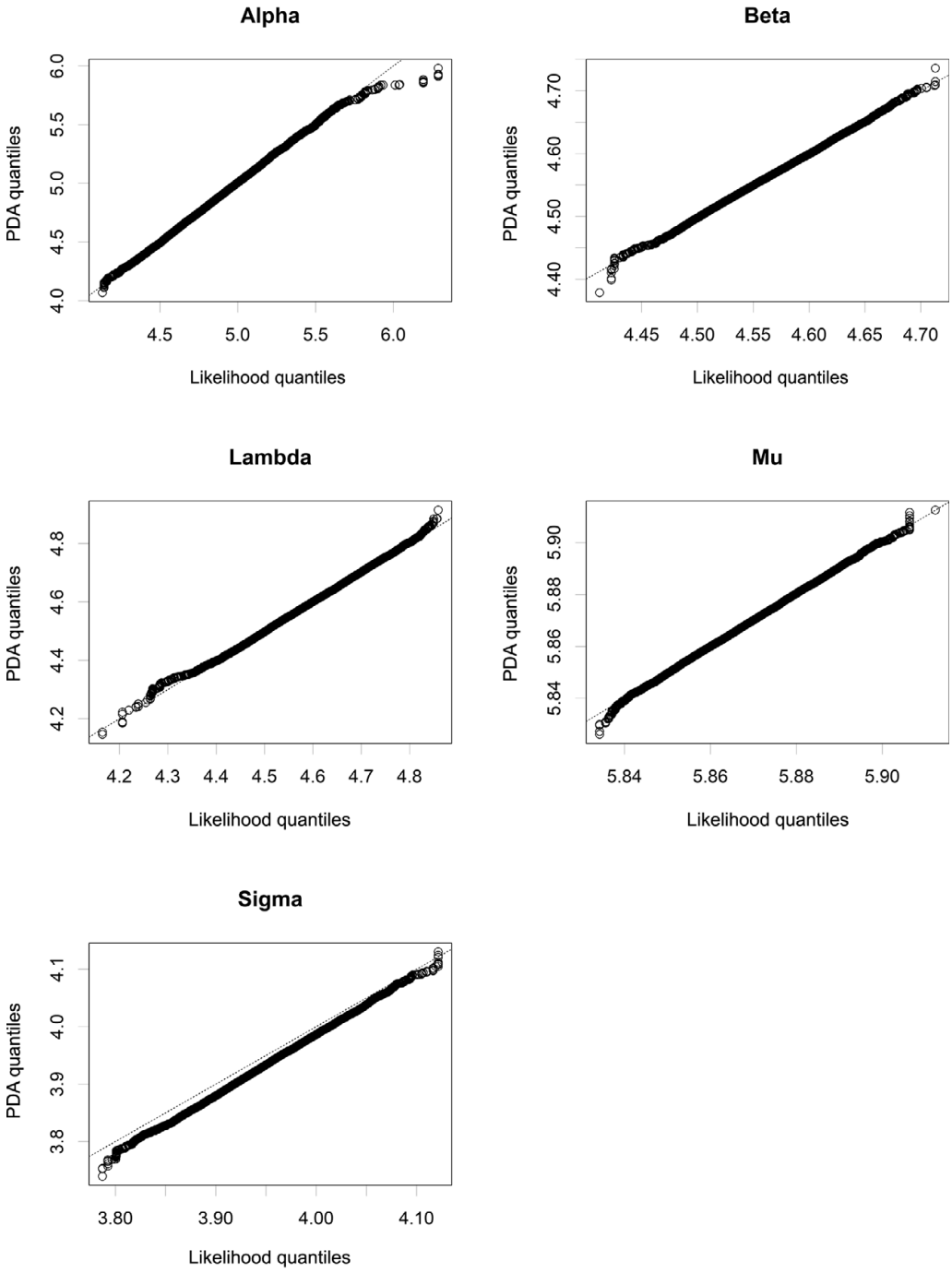


Figure 8.7 *Quantile–quantile plots of the estimated posterior distributions for the logged ASR parameters (α, β, λ_{inh}, μ_C, and σ, filled left to right, top to bottom), contrasting the likelihood-based MCMC and probability density approximation algorithms. The dotted line is the identity.*

The PMC method we demonstrated in this chapter is optimal in the sense that it results in an estimated posterior distribution that is as close (in the Kullback–Leibler sense) to the proposal distribution, giving high acceptance rates. However, the method is highly dependent on the distance metric chosen and the sufficiency of the statistics that determine the distance. Our implementation of the PMC method produced highly inaccurate estimates despite making reasonable choices, and stands as an example of how an ABC algorithm can unexpectedly fail (see also Jaynes, 2003).

Finally, PDA can produce quite accurate estimates of the posterior distributions without as many choices to be made about distance metrics, tolerance criteria, and so forth. Because a distance does not need to be computed for this algorithm, the issue of sufficient statistics is moot. This is in contrast to the rejection and particle filtering algorithms where sufficient statistics are crucial.

While the selection of a particular estimation method will depend on the application involved and the programming skill of the researcher, the PDA method is the most reliable of those we have investigated in this chapter. It must be noted that there is no guarantee of estimation accuracy for any of the methods presented here (including standard likelihood-informed MCMC methods).

For additional resources, interested readers should consult the number of references describing ABC methods, including Palestro *et al.* (2018). All code used to generate the estimates discussed in this chapter may be found at https://github.com/noahmthomas-nmt/ABC_Chapter.

Acknowledgments

Completion of this chapter was supported in part by award number BCS-1847603 from the National Science Foundation to Brandon M. Turner. This material is based upon work performed while Trisha Van Zandt was serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bassett, R., & Deride, J. (2019). Maximum a posteriori estimators as a limit of Bayes estimators. *Mathematical Programming*, *174*, 129–144.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 379–406.

Beaumont, M. A., Zhang, W., & Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*, 2025–2035.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Berlin: Springer.

Busemeyer, J., & Townsend, J. (1993). Decision Field Theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459.

Cappé, O., Godsill, S. J., & Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, *95*, 899–924.

Cappé, O., Guillin, A., Marin, J. M., & Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, *13*, 907–929.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*, 30055–30062.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*, 410–418.

Didelot, X., Everitt, R. G., Johansen, A. M., & Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, *6*, 49–76.

Doob, J. L. (1942). The Brownian movement and stochastic equations. *Annals of Mathematics*, *43*, 351–369.

Douc, R., Guillin, A., Marin, J.-M., & Robert, C. P. (2007). Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, *35*, 420–448.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Eriksen, C. W., & Hoffman, J. E. (1974). Selective attention: Noise suppression or signal enhancement? *Bulletin of the Psychonomic Society*, *4*, 587–589.

Eriksen, C. W., & Schultz, D. W. (1977). Retinal locus and acuity in visual information processing. *Bulletin of the Psychonomic Society*, *9*, 81–84.

Eriksen, C. W., & Schultz, D. W. (1978). Temporal factors in visual information processing. In J. Requin (Ed.), *Attention and Performance VII.* New York: Academic Press.

Farrell, S., & Lewandowski, S. (2018). *Computational modeling of cognition and behavior*. Cambridge: Cambridge University Press.

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*, 419–474.

Fu, Y.-X., & Li, W.-H. (1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution*, *14*, 195–199.

Gelfand, A. E., & Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, *94*, 247–253.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*, 3rd ed. Boca Raton, FL: CRC Press.

Gordon, N., Salmond, D., & Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, *140*, 107–113.

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*, 772–789.

Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, *91*, 1461–1473.

Jaynes, E. T. (2003). *Probability theory: The logic of science* (G. L. Bretthorst, Ed.). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511790423

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. I). New York: Academic Press.

Liu, Q., Petrov, A. A., Lu, Z.-L., & Turner, B. M. (2020). Extensions of multivariate dynamical systems to simultaneously explain neural and behavioral data. *Computational Brain & Behavior*, *3*, 430–457.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software*, *42*(9), 22. doi: 10.18637/jss.v042.i09

Miller, J., & Schwarz, W. (2021). Delta plots for conflict tasks: An activation-suppression race model. *Psychonomic Bulletin & Review*, *28*, 1755–1775.

Palestro, J., Sederberg, P., Osth, A., Van Zandt, T., & Turner, B. (2018). *Likelihood-free methods for cognitive science*. Cham: Springer International Publishing.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*, 1791–1798.

Robert, C., & Casella, G. (2014). *Monte Carlo statistical methods*. New York: Springer.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, *7*, 110–120.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2017). Bayesian hierarchical models of cognition. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology: Foundations and methodology* (pp. 504–551). Cambridge: Cambridge University Press.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151–1172.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*, 300–304.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, *145*, 505–518.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for parameter estimation. *Psychonomic Bulletin & Review*, *21*, 227–250.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*, 69–85.

Turner, B. M., & Van Zandt, T. (2014). Hierarchical approximate Bayesian computation. *Psychometrika*, *79*, 185–209.

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148–174.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

Weiss, G., & von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics*, *149*, 1539–1546.

Wilkinson, R. D. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, *12*, 129–141.

9 Cognitive Diagnosis Models

Jimmy de la Torre and Miguel A. Sorrel

9.1 Introduction

In recent years, a family of psychometric models has been developed for classifying examinees against a number of discrete attributes. In this context, attributes are construed as latent categorical variables that can refer to skills, competencies, tasks, or cognitive processes, among others. These models are ideal in situations where the primary goal of assessment is to identify or classify examinees' statuses with respect to a set of attributes. Mathematically, the attribute of examinee i is represented by $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}'$, where K is the number of attributes of interest. The most common of these models leads to dichotomous classifications, as in $\alpha_{ik} \in \{0, 1\}$, to indicate whether or not the examinee has mastered attribute k.

This family of psychometric models has come to be known as *cognitive diagnostic models* (CDMs). These models emerge in response to a clamour to obtain formative and actionable feedback from test data, which can be difficult to achieve from traditional psychometric models designed to rank-order individuals on a

single continuum. Diagnosis is typically carried out as a first step in determining appropriate interventions or remedial actions. Specifically, in educational contexts, these models have been designed to give diagnostic feedback about students' weaknesses and strengths that thereafter can be used by teachers to inform their instruction or by students to direct their own learning. In this respect, CDMs represent a new psychometric framework to extract diagnostic information from test data.

This chapter aims to present the developments in the area in the past decade. In this sense, it intends to complement previous reviews carried out in the 2000s. The structure of the chapter bears some correspondence with the steps in the application of CDMs. First, a correspondence matrix is constructed between the items and the attributes being measured, which is evaluated theoretically and empirically (Sections 9.1.1 and 9.2). Second, a set of appropriate models is selected, according to absolute and relative fit information (Section 9.3). Third, the estimated parameters for the selected models are interpreted (Sections 9.1.1 and 9.1.2). Fourth, as with any psychometric model, evidence of reliability and the valid use of the estimated parameters is sought (Section 9.4). Finally, future trends in the area are discussed (Section 9.5).

9.1.1 Basic Ideas

The term "attribute" in the CDM literature is used analogously with ability in item response theory (IRT). As in conventional IRT, attributes in cognitive diagnosis modeling are construed as latent constructs and are represented by latent variables. Let the response vector of examinee i, $i = 1, 2, \ldots, I$ to J items be denoted by $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$. Like IRT models, CDMs require an $I \times J$ binary item response matrix $\boldsymbol{Y}$ as input; however, unlike most IRT models, CDMs additionally require a $J \times K$ binary Q-matrix. The rows of the Q-matrix pertain to the items, and the columns the attributes. The 1s in the jth row of the Q-matrix identify the attributes required for item j. This is similar to the cognitive model of test specifications that also uses a two-way matrix to establish content and skill groupings to obtain a representative sample of items during test construction from a defined achievement domain. Using such a model, test items are then generated to represent each combination of content and skill in the matrix. In this regard, the test specifications function as a cognitive model that reflects the knowledge and skills examinees are expected to use to answer test items correctly. These tests are typically designed to measure many different behaviors within a short time. Similarly, the Q-matrix must include a representative sample of items for the assessment to be able to generate the desired diagnostic information.

Table 9.1 gives the Q-matrix for three items similar to those in the often-used fraction-subtraction data. For illustration purposes, we examine the test specification for item 2. To be able to solve the problem $\frac{5}{8} - \frac{3}{8} =?$, students must know how to subtract basic fractions (attribute 1) and reduce the result to the simplest form (attribute 2). Thus, the substantive model requires that students

Table 9.1 *Q-matrix for three fraction-subtraction items.*

Item	Problem	Correct Response	Attribute 1	Attribute 2	Attribute 3
1	$\frac{6}{8} =?$	$\frac{3}{4}$	0	1	0
2	$\frac{5}{8} - \frac{3}{8} =?$	$\frac{1}{4}$	1	1	0
3	$4\frac{1}{4} - 2\frac{3}{4} =?$	$1\frac{1}{2}$	1	1	1

Note: The attributes are 1 = performing basic fraction-subtraction operation, 2 = simplifying/reducing, 3 = converting mixed numbers to fractions.

possess two specific attributes to have a high probability of answering the problem correctly.

In the unrestricted case, a total of $L = 2^K$ latent classes can result from K dichotomous attributes. The items in Table 9.1 measure $K = 3$ dichotomous attributes, thus, $L = 2^3 = 8$. The different CDMs express the conditional success probability on item j given the latent class $\boldsymbol{\alpha}_l$, as in $P(Y_j = 1|\boldsymbol{\alpha}_l)$, which we can write as $P(\boldsymbol{\alpha}_l)$ when there is no confusion. More often than not, an item measures only a subset of the attributes. Accordingly, $\boldsymbol{\alpha}_l$ can be simplified by collapsing across irrelevant attributes such that the resulting latent groups have homogeneous within-group success probabilities. Let K_j^* be the number of attributes measured by item j. Additionally, for notational convenience, we will assume that the first K_j^* attributes are required. The associated collapsed attribute vector can be denoted by $\boldsymbol{\alpha}_{jl}^*$, and, in the most general case, it can differentiate between $2^{K_j^*}$ latent groups. Without any constraints, a unique success probability (i.e., probability that cannot be derived from other probabilities) is associated with each latent group. A CDM is considered saturated when the number of parameters equals the number of latent groups.

In addition to saturated or general models, there exist several reduced or specific CDMs in the literature. Different classification schemes have been used to differentiate the different CDMs. Models are said to be conjunctive (disjunctive) if all (one or more) of the required attributes are necessary to answer the item successfully. Alternatively, models are said to be compensatory (noncompensatory) if the absence of a required attribute can (cannot) be made up for by the presence of other attributes. For the most part, these two CDM classification schemes have been used interchangeably. Specifically, conjunctive models are also deemed noncompensatory, and disjunctive models compensatory. Note, however, that depending on how the terms are defined, the two classification schemes may

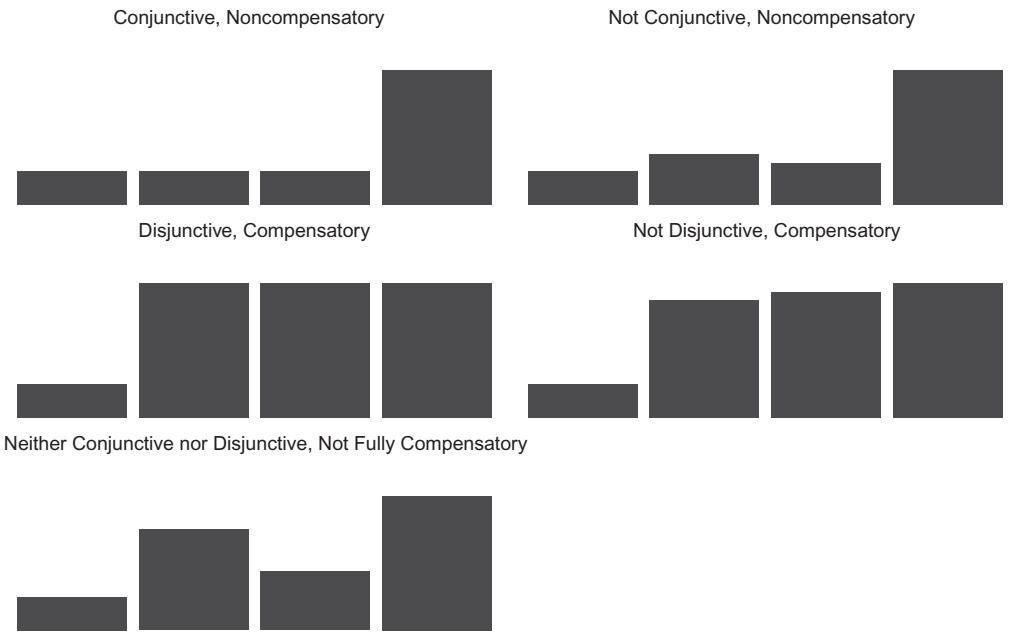


Figure 9.1 *Representation of the different possible CDM types that can be formed considering the conjunctive–disjunctive and compensatory–noncompensatory dimensions. From left to right, the x-axis indicates the latent groups* $\boldsymbol{\alpha}_{jl}^{*}$ = (00), (10), (01), *and* (11)*; the y-axis represents the probability of success for examinees in the latent groups represented on the x-axis.*

not be identical. Figure 9.1 represents the different possible combinations of these classification schemes.

One example of a conjunctive and noncompensatory CDM is the deterministic input, noisy AND gate (DINA) model. This is the most popular CDM in empirical applications according to a recent review, possibly because of its simplicity. The DINA model only differentiates between two latent groups for item j – those examinees who mastered all the required attributes, ($\eta_{ij} = I[\boldsymbol{\alpha}_i'\boldsymbol{q}_j = \boldsymbol{q}_j'\boldsymbol{q}_j] = 1$), and those lacking at least one of them, ($\eta_{ij} = I[\boldsymbol{\alpha}_i'\boldsymbol{q}_j = \boldsymbol{q}_j'\boldsymbol{q}_j] = 0$). The model has two parameters per item, guessing and slip, where the former is defined as $g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$, and the latter as $s_j = P(Y_{ij} = 0|\eta_{ij} = 1)$. Therefore, the item response function (IRF) of the model can be written as

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = P(\boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}}. \tag{9.1}$$

The complement of the DINA model is the deterministic input, noisy OR gate (DINO) model, which is a disjunctive and compensatory model. The formulation of this model is very similar to that of the DINA model, with the peculiarity that the DINO model differentiates only those who mastered at least one of the required attributes, ($\eta_{ij}^{*} = I[\boldsymbol{\alpha}_i'\boldsymbol{q}_j \geq 0] = 1$), from those who did not master any,

($\eta_{ij}^* = I[\boldsymbol{\alpha}_i' \boldsymbol{q}_j \geq 0] = 0$). Aside from the straightforward interpretation, the two models have been widely studied because of their relative simplicity – the number of parameters per item is always two regardless of the number of attributes that the item measures, making these models relatively easy to estimate. However, it should be borne in mind that these models invoke very strong assumptions about the underlying process. For example, in the DINA model for an item measuring three attributes, the latent groups $\boldsymbol{\alpha}_{lj}^* = (000)$, (100), (001), (001), (110), (101), and (011) are assumed to have the same success probability. That is, examinees who have mastered none, one, or two out of the three required attributes are considered indistinguishable. In some contexts (e.g., having partial knowledge is better than no knowledge), this assumption may not be reasonable. To increase the generalizability, and hence applicability, of CDMs, models with less stringent assumptions have been proposed.

One such model is the generalized DINA (G-DINA) model, which can be viewed as a generalization of the DINA model. Instead of only two latent groups, the G-DINA model partitions the latent classes into $2^{K_j^*}$ latent groups. Each latent group represents one reduced attribute vector $\boldsymbol{\alpha}_{lj}^*$ and has its own associated success probability, denoted as $P(\boldsymbol{\alpha}_{jl}^*)$. For the identity link, the success probability under the G-DINA model is written as

$$P(\boldsymbol{\alpha}_{jl}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12+\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \tag{9.2}$$

where δ_{j0} is the intercept (baseline probability), δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$. The G-DINA model has $2^{K_j^*}$ parameters for item j.

In its unconstrained form, the G-DINA model is equivalent to other general CDMs (e.g., the general diagnostic model; the loglinear cognitive diagnosis model). The G-DINA model can also be expressed using the logit and log links. The log-odds (also, log-linear) CDM is

$$logit[P(\boldsymbol{\alpha}_{jl}^*)] = \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \lambda_{jkk}\alpha_{lk}\alpha_{lk'} + \cdots + \lambda_{j12+c\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}; \tag{9.3}$$

and the log CDM is

$$log[P(\boldsymbol{\alpha}_{jl}^*)] = v_{j0} + \sum_{k=1}^{K_j^*} v_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} v_{jkk}\alpha_{lk}\alpha_{lk'} + \cdots + v_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}. \tag{9.4}$$

Several commonly encountered CDMs, including the above-mentioned DINA and DINO models, can be shown to be special cases of the G-DINA model. When

all the coefficients in Equation (9.2) except for δ_0 and $\delta_{j12\ldots K_j^*}$ are set to zero, the G-DINA model reduces to the DINA model. The DINO model can be written as

$$P(\boldsymbol{\alpha}_{jl}^*) = \begin{cases} \delta_{j0}, & \text{if } \boldsymbol{\alpha}_{jl}^* = \mathbf{0}_{K_j*} \\ \delta_{j0} + \delta_{jk}, & \text{otherwise} \end{cases} . \tag{9.5}$$

This is the G-DINA model with the constraint $\delta_{jk} = -\delta_{jk'k''} = \cdots = (-1)^{(K_j^*+1)}\delta_{j12\cdots K_j^*}$. Finally, when all the interaction terms are dropped, the G-DINA model in the identity link reduces to the additive CDM (*A*-CDM; de la Torre, 2011), which indicates that mastering attribute k increases the success probability on item j by δ_{jk}, independent of the contributions of other attributes. Thus, this model is neither disjunctive nor fully compensatory. The *A*-CDM has $K_j^* + 1$ parameters for item j. The linear logistic model (LLM) and the reduced reparameterized unified model (RRUM) are also additive models under the logit and log links, respectively. Incidentally, models with additive nature can be viewed as another scheme of classifying CDMs.

The G-DINA model, which includes the reduced CDMs it subsumes, offers a framework for dichotomous attributes in conjunction with dichotomous data. However, other CDMs in the same vein that cover different attribute and response types exist. Here we cover four of them: CDMs for polytomous attributes, multiple-choice (MC) data, continuous data, and ordinal and nominal data. First, the polytomous G-DINA (pG-DINA) model is an extension of the G-DINA model to accommodate polytomous attributes. Its corresponding Q-matrix has been extended as follows: $q_{jk} = 0, 1, \ldots, M_k - 1$, where M_k represents the number of levels of α_k. When $q_{jk} > 0$, α_k is required for item j, and its value represents the minimum attribute level needed to answer the item correctly. By invoking the specific attribute level mastery (SALM) assumption, the pG-DINA model dichotomizes the polytomous attributes required for the items, as in $\alpha_k^{**} = I[\alpha_k \geq q_{jk}]$, for $q_{jk} > 0$ and $k = 1, \ldots, K_j^*$. By reducing the number of latent groups that can be formed and allowing the G-DINA model to be used with the dichotomized reduced attribute vector $\boldsymbol{\alpha}^{**}$, the SALM assumption facilitates the estimation of the pG-DINA model. In its current formulation, the required attribute levels are specified *a priori* by subject-matter experts.

The second extension takes into account the fact that, in many instances, dichotomous data arise from multiple-choice tests when responses are coded only as either correct or incorrect. This coding procedure can limit the diagnostic utility of MC tests because it ignores information that can be found in the distractors, which may be useful in further differentiating certain latent classes. To maximize the diagnostic value of MC tests, the MC-DINA model has been proposed. In this setup, the correct option and some distractors are coded (i.e., designed to correspond to some latent classes); the remaining distractors are deemed noncoded. The model assumes that coded distractors measure a subset of the attributes measured by the key. Let H_j and H_j^* represent the number of options and number of coded options of item j, respectively. Based on the H_j^* coded options, the 2^K

latent classes are partitioned into $H_j^* + 1$ latent groups, where the additional group consists of latent classes that do not correspond to any of the coded options. The MC-DINA model specifies the conditional probability of a latent group choosing option h, $h = 1, \ldots, H_j$. Thus, the number of parameters equals $(H_j^* + 1) \times (H_j - 1)$, which grows with the number of options, as well as the number of coded options. A variant of the MC-DINA model that substantially reduces the number of parameters under certain assumptions has also been proposed.

Third, another source of information or response type is response time, which nowadays is more readily available due to the proliferation of computer-based assessments. One CDM that can be used with continuous response is the continuous DINA (cDINA) model. Like the DINA model, the cDINA model partitions the examinees into two latent groups – examinees who have all the required attributes for an item and those who do not. However, instead of the slip and guessing parameters, the cDINA models the log-response time distribution of a latent group for each item. Specifically, the cDINA model estimates the mean ($\mu_{j\eta}$) and variance ($\sigma_{j\eta}^2$) of the log-response time distribution for latent group η. To address the strong assumptions of the cDINA model, the continuous G-DINA model, which can accommodate $2^{K_j^*}$ latent groups, was introduced. With more latent groups come additional parameters to be estimated, as in from 2 to $2 \times 2^{K_j^*}$.

Fourth, a final example of CDM extension allows these models to be used with constructed-response items. This type of item is typically scored polytomously, yielding graded response data with ordered categories. The sequential G-DINA model is designed to handle multi-category response, where the required attributes may vary across the categories. Thus, instead of a single item–attribute association, the model requires multiple category–attribute associations for an item. In addition to the category–attribute association feature, the model assumes that solving an item consists of a number of sequential steps, and an examinee's scores are based on the number of steps they have correctly answered. Table 9.2 provides examples of category–attribute associations for a fraction-subtraction item. The Q-matrix represented in Table 9.2 is said to be restricted because the attributes required for the different categories are not identical. An unrestricted version of the Q-matrix is involved when more than one category–attribute association is used across the different steps. The probability of an examinee with latent group $\boldsymbol{\alpha}_c$ answering category h of item j correctly, given that they have successfully completed category $h-1$, is called the processing function of category h, and denoted by $S_j(h|\boldsymbol{\alpha}_c)$. The probability of scoring h on item j for examinees with the attribute vector $\boldsymbol{\alpha}_c$ can be expressed as

$$P(Y_j = h|\boldsymbol{\alpha}_c) = [1 - S_j(h+1|\boldsymbol{\alpha}_c)] \prod_{y=0}^{h} S_j(y|\boldsymbol{\alpha}_c) \tag{9.6}$$

subject to the constraint that $\sum_{h=0}^{H_j} P(Y_j = h|\boldsymbol{\alpha}_c) = 1$. In this flexible formulation, any dichotomous CDM can be used as the processing function, and a different CDM can be associated with each category. Moreover, the unrestricted sequential

Table 9.2 *Sequential steps in solving a fraction-subtraction item.*

		Attribute		
Step	Category	1	2	3
$1\frac{1}{4} - \frac{3}{4}$	0	0	0	0
$\frac{5}{4} - \frac{3}{4}$	1	0	0	1
$\frac{2}{4}$	2	1	0	0
$\frac{1}{2}$	3	0	1	0

Note: The attributes are 1 = performing basic fraction-subtraction operation; 2 = simplifying/reducing, 3 = converting mixed numbers to fractions.

G-DINA model can be used with nominal-response data, and can be shown to be equivalent to the nominal-response diagnostic model and the partial-credit DINA model. In addition to model extensions, other recent developments include a general framework for polytomous responses, a diagnostic tree model for polytomous responses and multiple strategies, and a model that combines attribute classification and misconceptions.

9.1.2 Model Estimation

The complete CDM formulation requires the specification of the distribution of the attribute vector $\boldsymbol{\alpha}$. Let $p(\boldsymbol{\alpha}_l)$ denote the joint distribution of the attributes, $P(Y_j = y_j|\boldsymbol{\alpha}_l)$ the conditional probability of response (i.e., the CDM), and $P(Y_j = y_j)$ the marginal probability of response. There are different approaches to model the joint distribution of the attributes. The first option is to use the saturated model, which involves the 2^K possible $\boldsymbol{\alpha}$ vectors, and requires $2^K - 1$ parameters to be estimated. Consequently, the number of parameters of the saturated model grows exponentially with K. When K is relatively large, say greater than 15, implementation can be extremely slow, if not computationally problematic. Another approach is to use a higher-order latent trait formulation to model the relationships among the attributes. A higher-order latent trait $\boldsymbol{\theta}$ is posited such that the components of $\boldsymbol{\alpha}$ are assumed to be independent conditional on $\boldsymbol{\theta}$. The higher-order model can be formulated as a linear logistic model, as in

$$P(\alpha_k = 1|\boldsymbol{\theta}) = \frac{\exp(\lambda_{0k} + \boldsymbol{\lambda}'_{1k}\boldsymbol{\theta})}{1 + \exp(\lambda_{0k} + \boldsymbol{\lambda}'_{1k}\boldsymbol{\theta})}, \tag{9.7}$$

which is the probability of mastering α_k given $\boldsymbol{\theta}$. The number of parameters in the higher-order model is linear in K. For example, when $\boldsymbol{\theta}$ is assumed to be

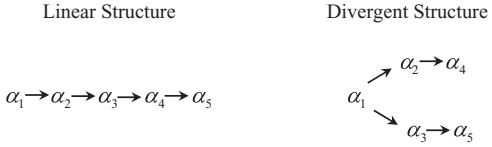


Figure 9.2 *Two general types of hierarchies that impose constraints on the attribute distribution.*

unidimensional, which is typically the case, the model has $2K$ parameters when a two-parameter logistic model is involved.

Alternatively, the attribute structure can be constrained based on a particular learning theory or curriculum that dictates the sequence by which the attributes are mastered. To illustrate this, Figure 9.2 gives two attribute structures for a test measuring five attributes. In the linear attribute structure, (mastery of) α_1 is a prerequisite to (mastery of) α_2, which in turn is a prerequisite to α_3, and so forth; in the divergent structure, α_1 is a prerequisite to α_2 and α_4, which in turn are prerequisites to α_3 and α_5, respectively. In this example, there are $2^5 = 32$ possible attribute vectors under the saturated attribute structure, and this number dramatically reduces to six and nine under the linear and divergent structures, respectively.

Markov chain Monte Carlo (MCMC) has been used with the higher-order models. In addition to the DINA model, other CDMs estimated with this model specification include the DINO model and the log-linear CDM. Although this estimation algorithm can easily be used even when the model is more complex or K is large, it can be computationally intensive. In contrast, the marginalized maximum likelihood (MML) estimation method has been used with saturated attribute distributions in conjunction with models such as the plain, continuous, and G-DINA models. When the CDMs are straight forward to estimate, MML estimation is generally very efficient up to moderate-sized K, but becomes inefficient as K gets larger. MML estimation has also been used when the attributes are constrained to be of a particular structure. On the one hand, constraining the attribute structure can lead to a more efficient estimation, but on the other hand, item parameter estimates and attribute classification can be very poor when an incorrect structure is used.

There are currently two general R packages that allow estimation of CDMs, as well as implementing other CDM-based methodologies (e.g., data simulation, Q-matrix validation, model-data fit assessment), namely, the `CDM` and the `GDINA` R packages. Other related R packages include `cdcatR` to conduct cognitive diagnosis computerized adaptive testing, `simcdm` to simulate CDM-based data simulation, and `NPCD` to implement nonparametric methods. The Bayesian estimation of the DINA model and the RRUM is possible through the `dina` and `rrum` packages. Although it is still commonplace for custom-built codes to be used in CDM research, the availability of these packages has made the implementation of these models greatly more accessible, particularly to applied researchers.

More general statistical software packages such as Mplus and the JAGS program can also be used to estimate CDMs, for which tutorials are available.

In the following, we illustrate the two estimation approaches with two examples. First, we discuss the MCMC estimation of the higher-order DINA model. In addition to the conditional distribution of $\boldsymbol{Y}$ given an attribute vector $\boldsymbol{\alpha}$ (i.e., a CDM), we also need the joint distribution of $\boldsymbol{\alpha}$. For the conditional distribution, we will use the DINA model; for the joint distribution specification, we will use the higher-order latent trait formulation. We will use MCMC to estimate the higher-order DINA model parameters. The IRF of the DINA model follows Equation (9.1), and the joint distribution of $\boldsymbol{\alpha}$ conditionally independent given unidimensional θ and with a common discrimination parameter can be formulated as

$$P(\boldsymbol{\alpha}|\theta,\boldsymbol{\lambda}) = \prod_{k=1}^{K} P(\alpha_k|\theta,\lambda_{0k},\lambda_1) = \prod_{k=1}^{K} \frac{\exp\left[1.7\lambda_1(\theta-\lambda_{0k})\right]}{1+\left[\exp 1.7\lambda_1(\theta-\lambda_{0k})\right]}. \tag{9.8}$$

The higher-order model lends itself to a hierarchical Bayesian formulation. To complete the model formulation, the prior distributions of $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, θ, $\boldsymbol{s}$, and $\boldsymbol{g}$ can be defined as

$$\begin{aligned}
\lambda_{0k} &\sim N(\mu_{\lambda_0},\sigma^2_{\lambda_0}),\\
\lambda_1 &\sim N(\mu_{\lambda_1},\sigma^2_{\lambda_1}),\\
\theta_i &\sim [N(\mu_\theta,\sigma^2_\theta),(\theta_i-\lambda_{0k})]^{-1},\\
\boldsymbol{\alpha}_{ik}|\theta_i,\lambda_{0k},\lambda_1 &\sim Ber(\{1+\exp[-1.7\lambda_1(\theta_i-\lambda_{0k})]\}^{-1}),\\
g_j &\sim 4\text{-}Beta(v_g,\omega_g,a_g,b_g), \text{ and}\\
s_j &\sim 4\text{-}Beta(v_s,\omega_s,a_s,b_s).
\end{aligned}$$

Invoking the conditional independence of $\boldsymbol{Y}$ given $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}$ given $\boldsymbol{\theta}$, the joint posterior distribution of $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, $\boldsymbol{s}$, and $\boldsymbol{g}$ given $\boldsymbol{Y}$ is

$$P(\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g}|\boldsymbol{Y}) \propto L(\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g}) \times P(\boldsymbol{\alpha}|\boldsymbol{\lambda},\boldsymbol{\theta}) \times P(\boldsymbol{\lambda}) \times P(\boldsymbol{\theta}) \times P(\boldsymbol{s}) \times P(\boldsymbol{g}),$$

where $L(\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g})$ is the likelihood of the data. Although the joint posterior distribution is complicated, it can be sampled using MCMC, especially Gibbs sampling. The full conditional distributions of $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, $\boldsymbol{s}$, and $\boldsymbol{g}$ are

$$\begin{aligned}
&P(\boldsymbol{\lambda}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g}) \propto P(\boldsymbol{\alpha}|\boldsymbol{\lambda},\boldsymbol{\theta})P(\boldsymbol{\lambda}),\\
&P(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g}) \propto P(\boldsymbol{\alpha}|\boldsymbol{\lambda},\boldsymbol{\theta})P(\boldsymbol{\theta}),\\
&P(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{s},\boldsymbol{g}) \propto L(\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g})P(\boldsymbol{\alpha}|\boldsymbol{\lambda},\boldsymbol{\theta}), \text{ and}\\
&P(\boldsymbol{s},\boldsymbol{g}|\boldsymbol{Y},\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{\alpha}) \propto L(\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{g})P(\boldsymbol{s})P(\boldsymbol{g}).
\end{aligned}$$

It can be noted that none of the full conditional distributions can be sampled directly. Hence, the Metropolis–Hasting method within Gibbs can be used with these distributions.

In the following we mainly cover the MML estimation of the DINA model with the saturated model for the joint distribution, where $L = 2^K$ possible attribute

vectors, $p(\boldsymbol{\alpha}_l)$, are considered. Assuming randomly sampled examinees and conditional independence of the responses given the attribute vector, the conditional likelihood of the observed data $\boldsymbol{Y}$ can be written as

$$\begin{aligned} L(\boldsymbol{Y} \mid \boldsymbol{\alpha}) &= \prod_{i=1}^{I} L(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_i) \\ &= \prod_{i=1}^{I} \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_i)^{Y_{ij}} [1 - P_j(\boldsymbol{\alpha}_i)^{1-Y_{ij}}]. \end{aligned} \tag{9.9}$$

The item parameters $\boldsymbol{\beta}$ and the attribute vectors $\boldsymbol{\alpha}$ can be simultaneously estimated using joint ML estimation (JMLE). However, as in traditional IRT, joint maximization of the structural parameter $\boldsymbol{\beta}$ and the incidental parameter $\boldsymbol{\alpha}$ can lead to inconsistent $\hat{\boldsymbol{\beta}}$. To arrive at consistent item parameter estimates, the latent variable can be integrated out of the conditional likelihood to obtain the marginalized likelihood of the data $\hat{\boldsymbol{\beta}}$ as follows:

$$L(\boldsymbol{Y}) = \prod_{i=1}^{I} L(\boldsymbol{Y}_i) = \prod_{i=1}^{I} \sum_{l=1}^{2^K} L(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l). \tag{9.10}$$

To obtain the maximum likelihood estimate of $\boldsymbol{\beta}_{j\eta}$, where $\beta_{j0} = g_j$ and $\beta_{j1} = s_j$, maximize

$$l(\boldsymbol{Y}) = \log \prod_{i=1}^{I} L(\boldsymbol{Y}_i) = \sum_{i=1}^{I} \log L(\boldsymbol{Y}_i) \tag{9.11}$$

with respect to $\boldsymbol{\beta}_{j\eta}$:

$$\frac{\partial l(\boldsymbol{Y})}{\partial \boldsymbol{\beta}_{j\eta}} = \sum_{i=1}^{N} \frac{1}{L(\boldsymbol{Y}_i)} \sum_{l=1}^{2^K} p(\boldsymbol{\alpha}_l) \frac{\partial L(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_l)}{\partial \boldsymbol{\beta}_{j\eta}}. \tag{9.12}$$

It can be shown that this maximization simplifies to solving for g_j and s_j in

$$\begin{aligned} \frac{1}{g_j(1-g_j)} [R_{jl}^{(0)} - g_j I_{jl}^{(0)}] = 0, \text{ and} \\ \frac{1}{(1-s_j)s_j} [R_{jl}^{(1)} - (1-s_j) I_{jl}^{(1)}] = 0, \end{aligned}$$

where $I_{jl}^{(0)}$ is the expected number of examinees lacking one or more of the attributes required for item j, and $R_{jl}^{(0)}$ is the expected number of examinees among $I_{jl}^{(0)}$ correctly answering the item; $I_{jl}^{(1)}$ and $R_{jl}^{(1)}$ carry the same interpretation, but they pertain to the examinees with all the attributes required for item j. Finally, the MML estimators are computed as $\hat{g}_j = R_{jl}^{(0)} / I_{jl}^{(0)}$ and $1 - \hat{s}_j = R_{jl}^{(1)} / I_{jl}^{(1)}$.

The same algorithm can easily be extended to estimate the G-DINA model parameters. The MML estimates of $P(\boldsymbol{\alpha}_{jl}^*)$ are

$$\hat{P}(\boldsymbol{\alpha}_{jl}^*) = \frac{\boldsymbol{R}_{\boldsymbol{\alpha}_{jl}^*}}{\boldsymbol{I}_{\boldsymbol{\alpha}_{jl}^*}} = \frac{\sum_{i=1}^{I} Y_{ij} p(\boldsymbol{\alpha}_{jl}^* | \boldsymbol{Y}_i)}{\sum_{i=1}^{I} p(\boldsymbol{\alpha}_{jl}^* | \boldsymbol{Y}_i)}, \tag{9.13}$$

where $p(\boldsymbol{\alpha}_{jl}^* | \boldsymbol{Y}_i)$ represents the posterior probability that examinee i is in latent group $\boldsymbol{\alpha}_{jl}^*$. The denominator represents the expected number of examinees in latent group $\boldsymbol{\alpha}_{jl}^*$, whereas the numerator represents the expected number of examinees in latent group $\boldsymbol{\alpha}_{jl}^*$ that will answer item j correctly.

There are different ways the standard errors for the item parameter estimates can be calculated. A simple way of computing the standard errors consists of taking the second derivative of $l(\boldsymbol{Y})$ with respect to $P(\boldsymbol{\alpha}_{jl}^*)$. It involves computing

$$\boldsymbol{I}[P(\boldsymbol{\alpha}_{jl}^*)] = -\sum_{i=1}^{I} \left\{ p(\boldsymbol{\alpha}_{jl}^* | \boldsymbol{Y}_i) \frac{Y_{ij} - P(\boldsymbol{\alpha}_{jl}^*)}{P(\boldsymbol{\alpha}_{jl}^*)[1 - P(\boldsymbol{\alpha}_{jl}^*)]} \right\} \left\{ p(\boldsymbol{\alpha}_{l'j}^* | \boldsymbol{Y}_i) \frac{Y_{ij} - P(\boldsymbol{\alpha}_{l'j}^*)}{P(\boldsymbol{\alpha}_{l'j}^*)[1 - P(\boldsymbol{\alpha}_{l'j}^*)]} \right\}. \tag{9.14}$$

Using the MML estimates to evaluate Equation (9.14), an approximation of the information matrix for the parameters of item j can be obtained (i.e., $\boldsymbol{I}[\hat{P}(\boldsymbol{\alpha}_{jl}^*)]$). The standard errors are found by taking the square root of the values on the main diagonal of $\boldsymbol{I}^{-1}[\hat{P}(\boldsymbol{\alpha}_{jl}^*)]$.

Without prior information, the joint attribute distribution can be initiated as $p(\boldsymbol{\alpha}_l) \sim \textit{Uniform}$. Thereafter, empirical Bayes' estimates can be obtained by updating the prior values of $p(\boldsymbol{\alpha}_l)$. Specifically, at iteration t, it can be updated as

$$p^{(t)}(\boldsymbol{\alpha}_l) = \frac{1}{I} \sum_{i=1}^{I} p^{(t-1)}(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i). \tag{9.15}$$

The estimation of reduced models can be carried out within the G-DINA model framework using design and weight matrices. Given that the DINA and DINO models have closed-form solutions, their parameters can be obtained as linear combinations of the G-DINA model parameter estimates. In contrast, additive models (i.e., A-CDM, LLM, RRUM) do not have closed-form solutions, hence, their parameters require optimization techniques to be obtained.

The MML algorithm described above has been modified to estimate other models. For example, to estimate CDMs with a hierarchical attribute structure, the prior probabilities of impermissible attribute vectors can be set to zero. Moreover, the G-DINA model can similarly be estimated by maximizing the derivative of Equation (9.11) with respect to the $2^{K_j^*}$ parameters. To obtain consistent item parameter estimates using JMLE, the procedure has been modified by incorporating a consistent estimator of the attribute patterns. To close this subsection, we briefly discuss two issues related to model estimation.

Model identifiability. One important topic related to model estimation is model identification, which is the set of minimum requirements for the model parameters

(e.g., item and person parameters) to be estimable from the observed data. In addition to the DINA and DINO models, the conditions for the identification for general models (e.g., G-DINA) have been established. For simpler models, it is necessary that the Q-matrix is complete. This implies that there is a single-attribute item measuring each attribute. This is a sufficient and necessary condition of identifiability of $p(\boldsymbol{\alpha}_l)$. With unknown $p(\boldsymbol{\alpha}_l)$ and $\boldsymbol{\beta}$, completeness is not enough to guarantee model identifiability. An additional requirement is that each attribute is measured by at least three items. In more recent work, different necessary and sufficient conditions have been provided for two-parameter and more general CDMs to be strictly or generically identified, which is consistent with the findings that a complete Q-matrix is not a prerequisite for less constrained CDMs to be identified.

Nonparametric methods. Depending on the models involved, reliable estimation of CDM parameters requires sufficiently large sizes. Such sample sizes may not be available in typical school settings, which can impede the CDM application where it is needed most. Nonparametric methods that bypass the estimation of model parameters to directly classify examinees have been developed to address this issue. These methods have been shown to provide more accurate attribute classification under very small sample size conditions (i.e., $N \leq 100$). Two nonparametric methods that have received some attention in recent years are the nonparametric classification (NPC) method and its generalization, the general NPC method. In the NPC method, the Hamming distance is used to compute the discrepancy between the observed and ideal response patterns following either a deterministic conjunctive (i.e., DINA-like) or disjunctive (i.e., DINO-like) rule. However, due to the restrictive nature of the DINA and DINO models, their fit to the data cannot always be guaranteed. To extend the practicability of the nonparametric methods, the GNPC was proposed. The GNPC, which is based on weighted ideal response patterns, is a more general method that can accommodate situations including and beyond the DINA and DINO models.

9.1.3 CDM Applications

It is not surprising that most CDM applications have been in the area of education, where these models first emerged. One of the first applications was in the domain of mixed-number subtraction. The same data set has been used by many researchers in the CDM context. A recent review of CDM applications found that 23 out of 74 papers focused on applied data analysis. A large majority of these papers (i.e., 86%) are in the areas of mathematics and reading, which indicates that CDM applications remain predominantly in the field of education. These applications include educational surveys for reading and mathematics assessments such as the National Assessment in Educational Progress (NAEP) and TIMSS; TOEFL and mock TOEFL; fraction arithmetic assessment; proportional reasoning; reading and listening comprehension; and spatial reasoning in the context of student learning.

Although scarce, applications in other areas also exist. Perhaps one of the most promising areas of CDM application is in diagnosing psychological disorders. The DINO model has been used to diagnose pathological gambling based on DSM-III, whereas the G-DINA model has been used to diagnose anxiety, somatoform, thought disorder, and major depression based on MCMI-III. In addition to existing measures, new instruments have been developed or validated from a CDM framework, and these include questionnaires to diagnose Internet gaming disorder based on DSM-V, and extroversion, neuroticism, callous unemotionality, and overt expressions of anger.

More recently, CDMs have also been applied in other domains. One of the first applications in the area of industrial-organizational psychology involves the use of CDMs to evaluate work competencies. In addition to an application of CDMs to measure entrepreneurial competencies, both dichotomous and polytomous CDMs have been used with situational judgment tests (SJTs) data.

A number of characteristics of empirical CDM applications have been documented. For example, the number of attributes these applications measured varied from four to 23, with four and eight being the mode and the mean, respectively. The sample size was greater than 1,000 in 61% of the studies examined. Finally, the most common CDMs were the DINA model and variations of the RUM, and approximately one-third of the studies estimated a general CDM, with the G-DINA model as the most frequently used model.

9.2 Q-Matrix Specification

Most, if not all CDMs, both general and specific, require a Q-matrix to identify the specific subset of attributes measured by each item. In most CDM applications, Q-matrix specification relies heavily on subject-matter or domain experts, and hence involves subjective judgments. Potential Q-matrix misspecifications resulting from the subjective nature of the Q-matrix construction process have raised serious validity concerns among researchers and practitioners. These misspecifications can degrade the quality of model parameter estimates, and, ultimately, the accuracy of the examinee attribute classifications. To minimize this problem, empirically based validity evidence must be gathered to examine the extent to which expert or theoretically based Q-matrix specifications are deemed acceptable. In the following subsection, we describe these two stages involved in constructing a Q-matrix.

9.2.1 Initial Q-Matrix Specification

Q-matrix construction is typically the first step in a CDM application. To this end, an initial list of attributes is drawn and the Q-matrix is specified based on these attributes. Prior research, relevant theories, expert rating tasks, and think-aloud protocols have been employed for these initial steps. In a prototypical expert

rating task, several domain experts are presented with the list of attributes and the corresponding operational definitions for their review and critique. To generate the initial Q-matrix, the experts are asked to identify the attribute/s required for each item. In some cases, the initial list of attributes or their definitions may be modified during this stage. To recognize the inherent uncertainty associated with these judgments, a modified coding scheme can be used. Specifically, q_{jk} can be coded as 1 (or 0) if it is certain that the attribute is required (or not required) and as 1^* if it is not clear whether the attribute is required or not. The Delphi method can then be implemented iteratively for several rounds. For example, one study involved three rounds, where experts identified the required attributes for each item in the first round, were anonymously provided with the results from the first round in the second round, and met in person to discuss the remaining discrepancies in the final round. To evaluate the degree of expert agreement in each round, the Fleiss' Kappa statistic was used. The above discussion assumes an extant assessment that can be used for diagnostic purposes. In situations where diagnostic assessments need to be built from scratch, the steps involved require a few modifications to accommodate developing new items that measure a wider range of attribute combinations.

After the initial or provisional Q-matrix has been determined, the next step is to assess its fit to the empirical data, once they become available, using procedures specifically designed for this purpose. Before discussing a method for validating a provisional Q-matrix, it should be noted that recent developments have opened the possibility for a fully exploratory approach, where the Q-matrix, and possibly the number of attributes, is directly estimated or learned from the data. Expert-defined Q-matrices, when available, can also be leveraged and used as priors in estimating the Q-matrix from the data. However, for meaningful results that conform to theoretical expectations, the same, if not greater, rigor and care need to be taken in developing the assessment and collecting the data.

9.2.2 Empirical Q-Matrix Validation

As noted in the previous subsection, the provisional Q-matrix may contain misspecifications that, if left unaddressed, can affect the valid use of the test scores. It is important to recognize that the Q-matrix is a component of the complete model specification. Thus, any model fit analysis should include the verification of the Q-matrix specifications. In recent years, various methods have been developed to assess this provisional Q-matrix based on the empirical evidence available. These methods have been called empirical Q-matrix validation methods.

The current literature includes various methods that use model fit information, hypothesis testing approaches, and nonparametric methods. Arguably, one of the more popular methods is that based on the general discrimination index (GDI) ς_j^2, which can be used in conjunction with the G-DINA model and the models it subsumes. Based on the rationale that appropriate q-vectors will yield latent groups that have homogeneous success probabilities, the index can be used to identify

and change incorrectly specified q-entries of item j. From among the appropriate q-vectors, the q-vector that leads to the highest variability of probabilities of success given the most parsimonious subset of attributes is deemed correct.

Given the specification $\boldsymbol{q}_l$, the associated ς^2_{jl} represents the (posterior) weighted variance of the probabilities of success of different latent groups, and is computed as

$$\varsigma^2_{jl} = \sum_{l'=1}^{2^{K^*_j}} p(\boldsymbol{\alpha}^*_{jl'})[P(\boldsymbol{\alpha}^*_{jl'}) - \bar{P}(\boldsymbol{\alpha}^*_j)]^2, \tag{9.16}$$

where $\boldsymbol{\alpha}^*_{jl'}$ is a $\boldsymbol{q}_l$-implied latent group, $p(\boldsymbol{\alpha}^*_{jl'})$ is the weight of the latent group l', and $\bar{P}(\boldsymbol{\alpha}^*_j) = \sum_{l'=1}^{K^*_j} p(\boldsymbol{\alpha}^*_{jl'})P(\boldsymbol{\alpha}^*_{jl'})$ is the mean success probability. To identify the correct q-vector, a q-vector with a particular ς^2_{jl} will be replaced by a q-vector that produces a higher $\varsigma^2_{jl'}$. Theoretically, all q-vectors that contain the required attributes for the item will achieve the maximum ς^2_{jl}, and a unique solution is arrived at by choosing the q-vector that excludes attributes that are irrelevant for the item.

When estimation error is involved, the highest variance is uniquely attained when $\boldsymbol{q}_{jL} = \mathbf{1}$, as in the saturated q-vector (i.e., all attributes are required) is specified. All the other possible $\boldsymbol{q}_{jl}$ are compared against the saturated q-vector by computing the proportion of variance accounted for, $PV\,AF_{jl} = \varsigma^2_{jl}/\varsigma^2_{jL}$. The suggested $\boldsymbol{q}_{jl^*}$ for item j is the most parsimonious q-vector with $PVAF_{jl} \geq \epsilon$. The performance of this method in terms of true positive rate and true negative rate for different cutoff values of ϵ was evaluated, and it was found that the optimal results can be obtained when the data conditions are considered in determining ϵ.

For the final q-vectors for item j to be deemed meaningful, the suggested $\boldsymbol{q}_{jl^*}$ needs to be judged based on their theoretical support. To reach this goal, the `GDINA` package includes a graphical tool called the mesaplot to facilitate in the decision-making process. The mesaplot represents the *PVAF* associated with each q-vector. For simplicity, only the q-vector with the highest *PVAF* is usually represented for each complexity group determined by $K^*_j = 1, \ldots, K$. Figure 9.3 provides an illustrative example for a simulated data set, where the true q-vector for item 10 $\boldsymbol{q}_{10} = (00001)$, was used in data generation and the overspecified q-vector, $\boldsymbol{q}_{10} = (00011)$, in data calibration. Using the default cutoff $\epsilon = 0.95$, all the q-vectors that include α_5 have $PVAF > \epsilon$. From among these q-vectors, (00001), the correct q-vector, is suggested because it contains the fewest specifications. Incidentally, the name "mesaplot" reflects the ideal condition where the incorrect q-vectors are separated from the appropriate q-vectors to form a mesa, and the correct q-vector sits at the edge of the mesa.

It should be noted that the procedure described above for arriving at the suggested q-vectors assumes that the provisional Q-matrix is true for the purpose of estimating the item parameters and the posterior distribution, which are the bases for computing ς^2_{jl}. Results based on a non-iterative validation procedure are

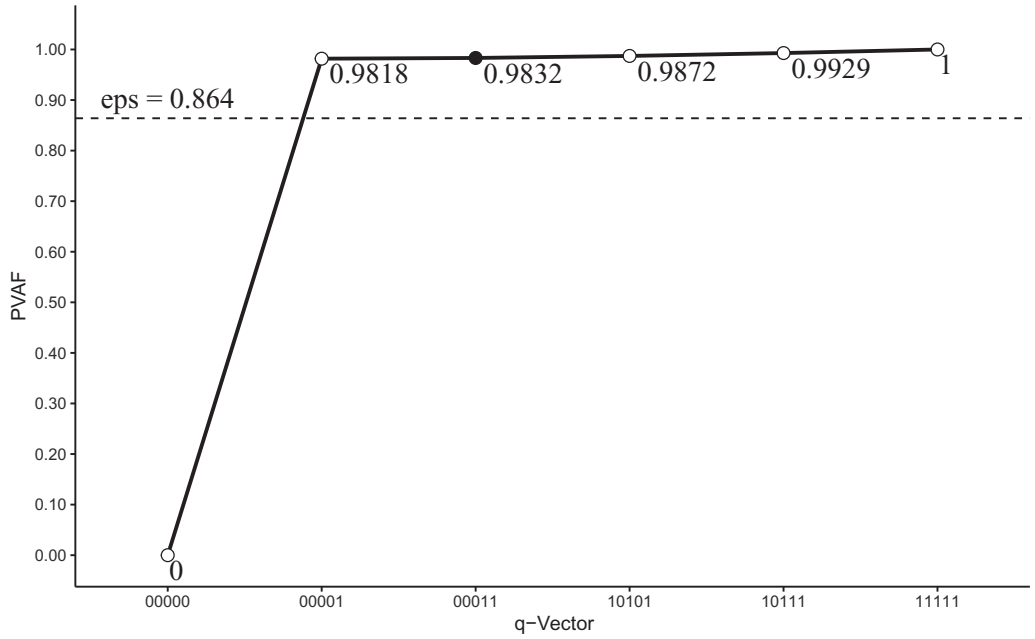


Figure 9.3 *Mesaplot for item 10:* $\boldsymbol{q}_{10} = (00001)$ *was used in the data generation and* $\boldsymbol{q}_{10} = (00011)$ *was used in the model calibration.*

suboptimal because of the contamination resulting from the misspecifications in the provisional Q-matrix. This issue has been addressed in the context of the GDI by employing an iterative procedure. At a particular iteration, only the suggested q-vector corresponding to the largest $\Delta PVAF$ is accepted. The validation process is repeated by recalibrating the model using the updated provisional Q-matrix, and terminated when no further suggestions are made. This iterative procedure has been shown to remain robust even with relatively large proportions of provisional Q-matrix misspecifications.

9.3 Model Fit Evaluation

Models are useful only to the extent they fit data. When different CDMs are available, they can be compared based on their fit to the data. Model comparison can be carried out at the item or test levels. At the item level, the residuals between the observed and expected moments, in particular the correlation and log-odds ratio between item pairs, can be compared. In addition to model selection, item-level residual analysis can be performed to evaluate the fit of a single model. Different CDMs can also be compared at the test level using the deviance (i.e., $-2LL$), the Akaike information criterion (*AIC*), and the Bayesian information criterion (*BIC*), as well as the Bayes factor and deviance information criterion (*DIC*) when the analysis is done using MCMC. These comparisons are

done without formally testing one model against another. Nevertheless, given that all reduced models are nested within the G-DINA model, the likelihood ratio test can be performed to statistically examine the adequacy of fit of a reduced model against that of the G-DINA model.

9.3.1 Absolute Fit

Absolute fit evaluation examines how well the model fits the data at hand. The most common absolute fit indices are residual-based. Appropriate models should result in estimates that predict essential characteristics of the data (e.g., inter-item correlations), thus producing small residuals. The most common way to assess absolute adjustment has been to assess item-level adjustment statistics. However, before discussing these statistics, we note that the $M2$, a test-level statistic, is an exception to this. It has been shown that $M2$, which compares residuals by item pairs and can be obtained from the GDINA R package, has adequate Type I error rates and statistical power. It also has a descriptive measure called RMSEA2, for which the cutoff points 0.045 and 0.030 have been suggested as indications of adequate and excellent fit, respectively.

At the item level, three residual-based statistics, namely, the proportion of correct individual items (p), the correlations (r), and the log-odds ratio of item pairs (l), have been introduced. In all three cases, the fitted model is used to simulate model-based item responses using a large generated sample size I^*. Of the three, r and l had very similar performance, and can detect CDM or Q-matrix misspecifications at a high rate. Due to their similarity, only the r statistic is discussed here. The observed and predicted response vectors for item j are indicated by the column vectors $\boldsymbol{Y}_j = \{Y_{1j}, \ldots, Y_{ij}, \ldots, Y_{Ij}\}'$ and $\tilde{\boldsymbol{Y}}_j = \{\tilde{Y}_{1j}, \ldots, \tilde{Y}_{ij}, \ldots, \tilde{Y}_{I^*j}\}'$, respectively. The r-statistic for items j and j' is computed as

$$r_{jj'} = \mid Z[Corr(\boldsymbol{Y}_j, \boldsymbol{Y}_{j'})] - Z[Corr(\tilde{\boldsymbol{Y}}_j, \tilde{\boldsymbol{Y}}_{j'})] \mid , \tag{9.17}$$

where $Corr(\cdot)$ is Pearson's product-moment correlation and $Z[\cdot]$ is the Fisher transformation. The approximate standard error of this statistic is given by $SE[r_{jj'}] = \sqrt{[I-3]}$. If the model is adequate for the data, this statistic is expected to equal zero for all items. Given the large number of comparisons involved (i.e., $J(J-1)/2$), the authors recommended examining only the largest statistic and adjusting the significance level using the Bonferroni correction. In Figure 9.4, heatmaps generated using the `GDINA` package are shown for the residuals from fitting the DINA and DINO models to DINA-generated data. It can be observed that multiple residuals are found to be significant in the incorrect model to indicate that the DINO model is not appropriate for these data, whereas the DINA model provides an acceptable fit.

Another popular measure of item fit is the root mean square error of approximation (*RMSEA*). The *RMSEA* for item j can be computed as

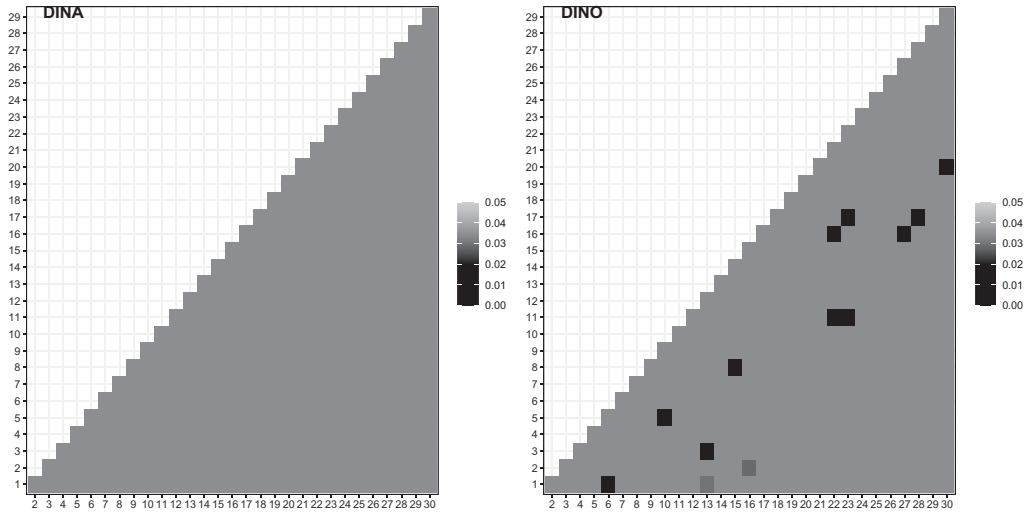


Figure 9.4 *Heatmaps of the item-pair correlation residuals for data generated using the DINA model. Cells in black indicate various degrees of significant residuals.*

$$RMSEA_j = \sqrt{\sum_{l=1}^{L} p(\boldsymbol{\alpha}_l)[P_{obs}(\boldsymbol{\alpha}_l) - P_{exp}(\boldsymbol{\alpha}_l)]^2}, \tag{9.18}$$

where $P_{exp}(\boldsymbol{\alpha}_l)$ and $P_{obs}(\boldsymbol{\alpha}_l)$ are the expected and observed success probabilities for examinees in latent class $\boldsymbol{\alpha}_l$, and $p(\boldsymbol{\alpha}_l)$ is the size of the latent class. The observed success probabilities are obtained using the estimated latent class memberships. The cutoffs 0.05 and 0.10 have been suggested as a general guideline to evaluate the size of the misfit. The $S - X^2$ statistic can be used to compare the expected and observed frequencies. This statistic is computed as

$$S - X_j^2 = \sum_{s=1}^{J-1} I_s \frac{(O_{js} - E_{js})^2}{E_{js}(1 - E_{js})}, \tag{9.19}$$

where s denotes an observed score group based on the sum scores, I_s is the number of examinees in group s, and O_{js} and E_{js} are the observed and predicted proportions of correct responses for item j. This statistic is assumed to be χ^2-distributed with $J - 1 - P$ degrees of freedom, where P is the number of item parameters. The model-predicted probabilities are computed as

$$P(y_{ij} = 1 \mid S_i = s) = \frac{\sum_{l=1}^{2^K} P(y_{ij} = 1 \mid \boldsymbol{\alpha}_l) P(S_i^j = s - 1 \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K} P(S_i = s \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l)}, \tag{9.20}$$

where $P(S_i^j = s - 1 \mid \boldsymbol{\alpha}_l)$ denotes the probability of obtaining the sum score $s - 1$ in the test composed of all items except item j. The two statistics above can be computed using the `CDM` package. A study has shown that the performance of the

item fit statistics can be expected to improve by adjusting for the measurement error in $\boldsymbol{\alpha}_l$; however, such adjustments have yet to be implemented in existing software packages.

9.3.2 Relative Fit

Relative to the saturated G-DINA model, reduced CDMs provide worse absolute fit to the data. To determine which of the models that provide good absolute fit to use, relative fit statistics can be used. These statistics are generally calculated from the maximum likelihood function that is obtained from the estimated ML parameters given in Equation (9.10).

As noted above, saturated CDMs (e.g., the G-DINA model), which have greater complexity, are theoretically expected to provide better fit to the data than reduced CDMs. However, saturated CDMs are not always to be preferred because they require a larger sample size to be well estimated. Moreover, reduced CDMs are simpler and easier to interpret, and when appropriate, lead to higher attribute classification accuracy. To choose between saturated and reduced CDMs, relative fit statistics that can compensate for model complexity are needed. Examples of these statistics are *AIC* and *BIC*, which can be computed as follows:

$$AIC = -2\log L(\boldsymbol{Y}) + 2P, \text{ and} \tag{9.21}$$

$$BIC = -2\log L(\boldsymbol{Y}) + P\log(I), \tag{9.22}$$

where lower values indicate a better balance between model-data fit and model complexity, and these statistics can be used for non-nested models. Results of a study examining the performance of these statistics generally supported the use of BIC, and to some extent, AIC, for evaluation of model or Q-matrix misspecifications. When nested models are involved (i.e., G-DINA model vs. reduced CDMs), formal tests can be performed to examine the adequacy of the fit of simpler models relative to that of a more complex model. Let S and R be the saturated and reduced models, respectively. The likelihood ratio (LR) statistic for comparing R and S is computed as

$$LR = 2[\log L^{(S)}(\boldsymbol{Y}) - \log L^{(R)}(\boldsymbol{Y})], \tag{9.23}$$

and is asymptotically χ^2-distributed with degrees of freedom equal to the difference of model parameters. In addition to the test-level comparison, saturated and reduced models can also be compared at the item level (i.e., one item at a time), and the LR test has been applied for this purpose. The direct implementation of the method would require estimating $J_{K_j^*>1} \times N_R + 1$ models, where N_R is the number of reduced models being considered. To allow for a more efficient implementation, the reduced model parameters can be estimated by maximizing the likelihood of the reduced parameters involved ($\boldsymbol{\psi}_j$) given $\boldsymbol{I}_j = \{I_{\boldsymbol{\alpha}_{jl}^*}\}$ and $\boldsymbol{R}_j = \{R_{\boldsymbol{\alpha}_{jl}^*}\}$, the G-DINA estimates of number of examinees and correct responses in the latent group $\boldsymbol{\alpha}_{jl}^*$, respectively. Recall that the ML estimator for the item parameters

in the saturated model is equal to $P(\boldsymbol{\alpha}_{jl}^*) = R_{\boldsymbol{\alpha}_{jl}^*}/I_{\boldsymbol{\alpha}_{jl}^*}$. This approximation has been referred to as the two-step likelihood ratio test, and is computed as

$$2LR_j = 2[\log L(\boldsymbol{P}_j \mid \boldsymbol{R}_j, \boldsymbol{I}_j) - \log L(\boldsymbol{\psi}_j \mid \boldsymbol{R}_j, \boldsymbol{I}_j)], \tag{9.24}$$

where $\boldsymbol{P}_j = \{P(\boldsymbol{\alpha}_{jl}^*)\}$. This approach requires that the data be calibrated once only using the G-DINA model; the remaining computations involve deriving $\boldsymbol{\psi}_j$ N_R times for the $J_{K_j^*>1}$ multi-attribute items.

The Wald statistic has also been introduced to compare saturated and reduced models at the item level. This statistic is computed as

$$W_j = [\boldsymbol{R} \times \boldsymbol{P}_j]'[\boldsymbol{R} \times Var(\boldsymbol{P}_j) \times \boldsymbol{R}']^{-1}[\boldsymbol{R} \times \boldsymbol{P}_j], \tag{9.25}$$

where $\boldsymbol{R}$ is a $(2^{K_j^*} - P) \times 2^{K_j^*}$ matrix of restrictions that make the reduced model a special case of the saturated model. For example, for the A-CDM and $K_j^* = 3$, $\boldsymbol{R}$ is equal to

$$\boldsymbol{R}_{4\times 8} = \begin{bmatrix} 1 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix}, \tag{9.26}$$

where each row represents a particular constraint and each column one of the eight latent groups that can be formed with $K_j^* = 3$. This restriction matrix implies the following constraints to the IRF of the G-DINA model in Equation (9.2): $\delta_{12} = \delta_{13} = \delta_{23} = \delta_{123} = 0$ (i.e., all the interaction terms are equal to 0). As with the item-level LR test, this method only requires a single calibration using the G-DINA model. In addition, estimates of the reduced models are not needed to compute the Wald statistic. Studies have shown that the two statistics produce acceptable Type I error and power rates.

9.4 Examinee Classification, Reliability, and Validity

9.4.1 Examinee Classification

As in IRT, the person parameter estimates in CDM (i.e., $\hat{\boldsymbol{\alpha}}_i$) can be based on ML estimation, maximum *a posteriori* (MAP), or expected *a posteriori* (EAP) methods. Recall that for examinee i, the likelihood $L(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_l)$ is defined as

$$L(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P(\boldsymbol{Y}_j = 1 \mid \boldsymbol{\alpha}_l)^{Y_{ij}}[1 - P(\boldsymbol{Y}_j = 1 \mid \boldsymbol{\alpha}_l)]^{1-Y_{ij}}. \tag{9.27}$$

The ML, MAP, and EAP estimators of $\boldsymbol{\alpha}_i$ are given by

$$ML(\boldsymbol{\alpha}_i) = \arg\max_{\boldsymbol{\alpha}_l}[L(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_l)], \tag{9.28}$$

$$MAP(\boldsymbol{\alpha}_i) = \arg\max_{\boldsymbol{\alpha}_l}[P(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i)], \text{ and} \tag{9.29}$$

$$EAP(\boldsymbol{\alpha}_i) = \{P(\alpha_{ik} \mid \boldsymbol{Y}_i)\} = \left\{ \sum_{l=1}^{L} P(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i)\alpha_{lk} \right\}, \tag{9.30}$$

where $P(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i)$ and $p(\boldsymbol{\alpha}_l)$ are the posterior and prior probabilities of $\boldsymbol{\alpha}_l$, respectively. It can be noted that $ML(\boldsymbol{\alpha}_i) = MAP(\boldsymbol{\alpha}_i)$ when $p(\boldsymbol{\alpha}_l)$ is flat, and $ML(\boldsymbol{\alpha}_i)$ and $MAP(\boldsymbol{\alpha}_i)$ are binary vectors, whereas $EAP(\boldsymbol{\alpha}_i)$ is a vector of probabilities. When classifying an examinee to one of the latent classes is of interest, the probabilities can be converted to 1s and 0s using certain rules (e.g., $\boldsymbol{\alpha}_{ik} = 1$ if $P(\boldsymbol{\alpha}_{ik} \mid \boldsymbol{Y}_i) \geq 0.50$). In some applications, a more stringent rule can be implemented such that a probability is converted to 0 (or 1) only when the $P(\boldsymbol{\alpha}_{ik} \mid \boldsymbol{Y}_i)$ is sufficiently small (or large); the remaining probability values (e.g., $[0.3, 0.7]$) comprise the uncertainty region, where no conversions are made. Finally, when $P(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i)$ is multimodal, for example, due to an incomplete Q-matrix, $ML(\boldsymbol{\alpha}_i)$ and $MAP(\boldsymbol{\alpha}_i)$ may not be unique.

Due to the effort expended in defining the attributes, CDM scores are generally interpretable. Assume that the CDM application takes place in the school classroom context. In addition to providing the students with their attribute profiles that will allow them to identify their individual strengths and weaknesses, the teacher might also be interested in obtaining information about the performance of the class as a whole to better determine how the instructional materials can be designed or scaffolded to target the specific needs of the class. Figure 9.5 shows an example of output that provides diagnostic information on three attributes at the student and classroom levels. The top panels of the figure display the attribute profiles of two students, and the bars and shades represent the EAP estimates for each attribute. The figure shows that student A has clearly mastered attributes 2 and 3, but the mastery status of attribute 1 is uncertain; in contrast, the panel shows that student B has clear mastery statuses for the three attributes – the student has definitely mastered attribute 3, but not attributes 1 and 2. The bottom left panel of the figure gives the percentage of students who have mastered each of the three attributes. At a glance, the teacher can easily note that attribute 3 has the highest mastery prevalence, whereas attribute 1 has the lowest. Perhaps subsequent instructions should focus on helping more students master attributes 1 and 2. Lastly, the bottom right panel disaggregates the three mastery prevalances into eight latent classes to better understand the prevalences of the different mastery profiles. It shows that $\boldsymbol{\alpha} = (0, 0, 1)$ and $\boldsymbol{\alpha} = (1, 0, 0)$ are the largest and smallest latent classes, respectively.

9.4.2 Reliability

The extent to which subsequent actions must be pursued may depend on how well the person parameters have been estimated. In CDM, reliability of the person parameter estimates, which is typically referred to as attribute classification accuracy, has been evaluated in many ways. A procedure for evaluating reliability is the Monte Carlo approach, which consists of the following steps:

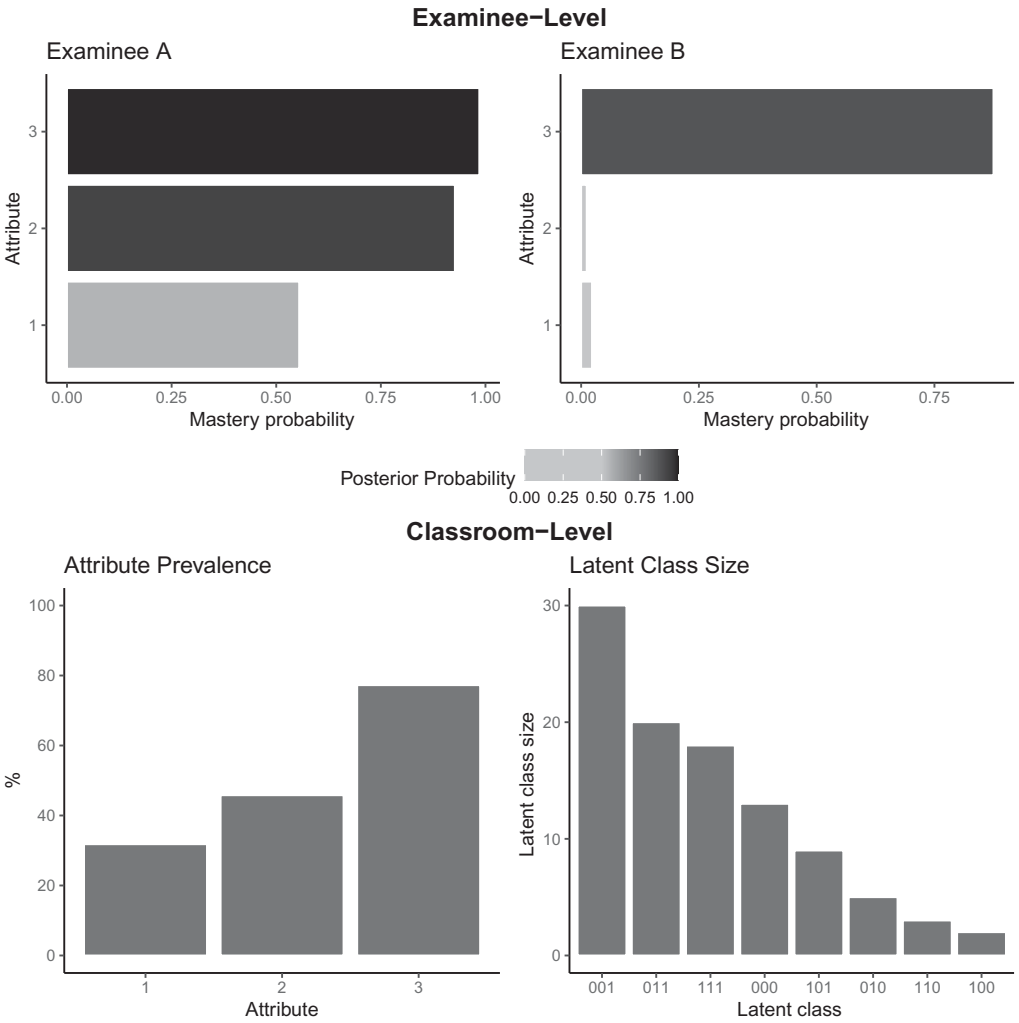


Figure 9.5 *Examples of CDM reports at the examinee and classroom levels*

1. First, the calibrated model (i.e., $\hat{\boldsymbol{\beta}}$) is used to generate a large simulated data set (e.g., $I^* = 10{,}000$). Ideally, this model is chosen after evaluating its fit to the data, as discussed in Sections 9.2 and 9.3.
2. Second, the same model is used to estimate the person parameters from the simulated data.
3. The agreement rates between the true and estimated person parameters are evaluated at the attribute vector and attribute levels by computing

$$PCV = \frac{1}{I}\sum_{i=1}^{I^*} I[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i], \text{ and} \tag{9.31}$$

$$PCA = \frac{1}{I}\sum_{i=1}^{I^*}\sum_{k=1}^{K} I[\alpha_{ik} = \hat{\alpha}_{ik}], \tag{9.32}$$

respectively, where I^* is the number of simulated examinees and $I[\cdot]$ is the indicator function. These two indices provide an estimate of the attribute classification accuracy in the empirical data set.

Other, more analytical procedures for computing reliability have also been proposed, and some of them have been implemented in the `GDINA` and `CDM` packages. One of the first methods, which is based on the $\hat{P}_a$ index, evaluates the classification accuracy at the test level. In contrast, the indices $\hat{\tau}$ and $\hat{\tau}_k$ have been introduced to evaluate test- and attribute-level reliabilities, respectively. Incidentally, $\hat{\tau}$-based indices require much simpler calculations compared to $\hat{P}_a$. The τ indices can be estimated from the examinees' posterior distributions [i.e., $P(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i)$], and this approach has two important advantages. Firstly, the calculations are very simple as they are obtained directly from information already available from the estimation process. And secondly, it provides information at the latent class level, which can then be marginalized to obtain indicators at the attribute ($\hat{\tau}_k$) and test $\hat{\tau}$ levels. This approach involves the calculation of a $2^K \times 2^K$ matrix of conditional classification error probabilities given by

$$P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{Y}) = \frac{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i) I[\boldsymbol{\alpha}_s = \boldsymbol{\alpha}_l]}{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i)}. \tag{9.33}$$

The main diagonal of $P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{Y})$ contains the classification accuracy estimator at the latent class level ($\hat{\tau}_l$). The sum of the 2^K elements contained in the main diagonal weighted by the estimated latent class proportions [i.e., $\hat{P}(\boldsymbol{\alpha}_l]$ results in $\hat{\tau}$. These values can also be marginalized for each individual attribute, obtaining an estimate for $\hat{\tau}_k$. The *CA*() function in the `GDINA` package is based on these developments. As an alternative, a new estimator of the classification accuracy, which is the basis for the *cdm.est.class.accuracy*() function in the `CDM` package, has been developed by extending the above indices. Given their similarities, the three procedures [i.e., the Monte Carlo approach, *CA*(), and *cdm.est.class.accuracy*()] are expected to yield similar values. As illustrated in Figure 9.6, the classification accuracy estimates at the attribute and attribute vector levels are highly comparable. Thus, all appear to be viable procedures. It should be noted that several other reliability indices exist.

9.4.3 Validity

As with any test scores, proper use of CDM scores requires not only ensuring adequate reliability, but also evidence of validity to be provided. Since the 1999 edition of the Standards for Educational and Psychological Testing, the validation process has been understood as a continuous process in which different types of evidence are sought. A large part of what has been said in Sections 9.2 and

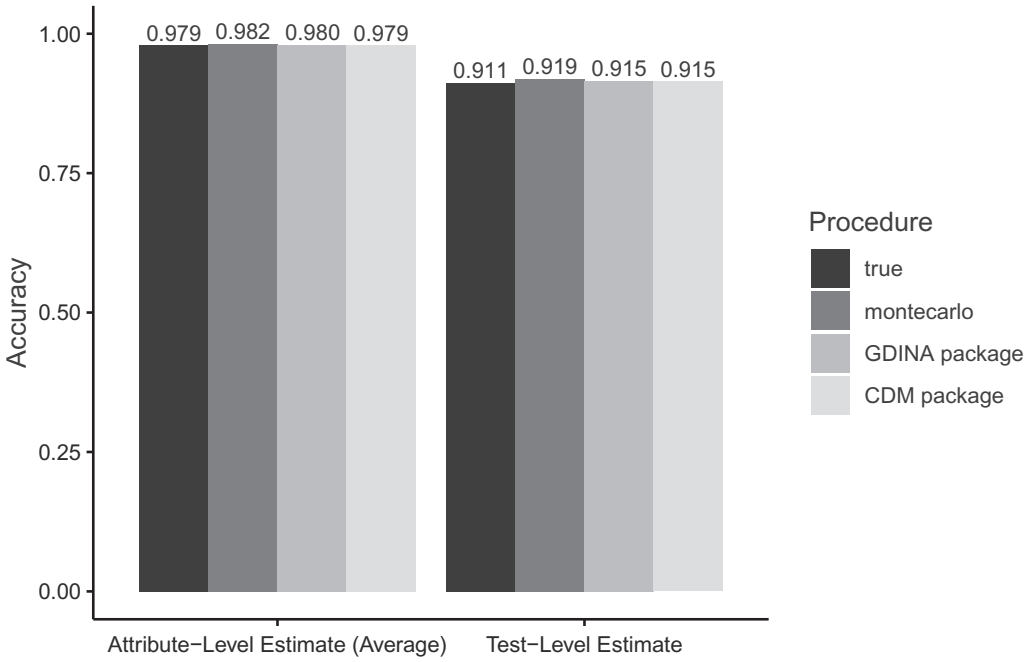


Figure 9.6 *Accuracy estimates for a test composed of 30 highly discriminating DINA items. true: True accuracy (i.e., PCA and PCV with respect to the generating attribute patterns). montecarlo: Monte Carlo PCA and PCV estimates based on a sample of 10,000 examinees. GDINA package: Estimates by the CA() function of the* `GDINA` *package. CDM package: Estimates by the cdm.est.class.accuracy() function of the* `CDM` *package.*

9.3 in relation to the evaluation of the Q-matrix and the selection of the CDM pertains to possible validity evidence of content, internal structure, and response processes. That is, the Q-matrix should be representative of the item population that represents the constructs it is intended to measure. It also establishes the relationships between these constructs (i.e., the attributes) and the different items. Finally, different CDMs reflect different response processes (e.g., compensatory or noncompensatory). Thus, information obtained from validating the Q-matrix and evaluating the model fit must be formulated in terms of a validity argument.

Another important aspect of CDM for which empirical support can be sought is determining the number of dimensions the test intends to measure. This issue has not been addressed in-depth, but seminal work has been started. Furthermore, to determine the distinctiveness of the different attributes being considered, correlations between the attributes need to be reported. Ideally, this should be done on the basis of previous hypotheses. A review of the empirical applications available to date found that most studies (72%) did not report these correlations. Of the remaining 28%, almost all reported correlations greater than 0.90, which is an artifact of fitting a multidimensional model (i.e., CDM) to largely unidimensional data.

In addition to the validity evidence discussed thus far, the importance of providing external validity evidence has been highlighted. For example, in examining an SJT that measures four attributes in a sample of university students, the grade-point average, scores on an advanced progressive matrices test, and NEO Five-Factor Inventory domain scores were used as criterion variables to validate the CDM scores. The results showed that the CDM scores obtained higher validity coefficients, compared to the SJT sum score. This is similar to a finding that, compared to traditional MCMI-III scores, CDM scores correlated more highly with a psychiatrist's diagnosis.

Finally, work has also been done on methods to obtain evidence pertaining to test fairness. In particular, methods to evaluate differential item functioning (DIF) have been developed. In the CDM context, an item is said to exhibit DIF when two examinees with identical attribute pattern (e.g., $\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_B$), but from different groups (e.g., gender, ethnicity) have different success probabilities on the item, as in $P(Y_j = 1 \mid \boldsymbol{\alpha}_A) \neq P(Y_i = 1 \mid \boldsymbol{\alpha}_B)$. One of the first works to address this issue in the CDM context involved the adaptation of the Mantel–Haenszel (MH) procedure, where the estimated latent class was used as the conditioning variable. Later work includes the use of the Wald test as a DIF detection procedure and formulating the problem from a Bayesian framework. A recent review noted that a number of scenarios can give rise to DIF. In particular, DIF can occur when: (1) the item parameters vary across groups; (2) different Q-matrices are involved; and (3) the underlying processes differ across groups. However, existing studies thus far focused only on the first scenario. An exhaustive simulation study found that using the MH statistic in conjunction with a purification procedure produced satisfactory Type I error and statistical power across the various DIF scenarios.

9.5 Discussion and Future Directions

This chapter attempts to summarize the main developments in CDMs over the last decade. However, due to space constraints it is not possible to include all the developments that have taken place during this period of time. For this reason, other developments – such as models for multiple strategies, testlets, and multi-level analysis – were not covered. This is undoubtedly a very active area of research, where interesting new work can be expected to emerge on a regular basis. Thus, in this final section we would like to highlight three lines of ongoing research that may attract the attention of more researchers in the near future. Below is a brief description of these research lines.

1. *Cognitive diagnostic computerized adaptive testing (CD-CAT)*. CD-CAT seeks to combine the specific feedback from CDMs with the efficiency of adaptive applications. In recent years, a number of developments related to item selection rules and optimal criteria, stopping rules and control of exposure, and content balancing have emerged. However, real applications and further studies examining other aspects of CD-CAT (e.g., designs to optimize the initial calibration

of the item bank) have been lagging behind. A recent study that partially fills this gap sought to explore the impact of specifying only a subset of the possible q-vectors. Related topics, such as optimal test assembly and multistage adaptive testing, have received much less attention beyond the few studies that currently exist. Given the critical role of CAT in facilitating the implementation of CDMs, it is not difficult to see that this area will continue to develop in the coming years.

2. *Exploratory CDMs*. The parametric models discussed in the chapter require the availability of a completely or partially correct provisional Q-matrix. Very recently, models that do not require a provisional Q-matrix have been developed. These models have come to be called exploratory CDMs, and are particularly useful in situations where there is no theoretical evidence or resources are limited to establish a provisional Q-matrix. This line of research takes the conditions for identifiability of the DINA model as the starting point. At present, models for dichotomous data (e.g., the exploratory DINA model, exploratory RRUM) and ordinal data are available. When warranted, these models are sufficiently flexible to allow for some elements of the Q-matrix to be fixed. Moreover, the number of attributes can also be learned from the data. Given the recent results on the identifiability conditions that apply to a wider range of CDMs, more general exploratory CDMs may be on the horizon.
3. *Measurement of learning*. With the school setting as the prototypical application context of CDMs, one vital function of these models is to measure and track learning. One of the first studies to address this issue in the context of CDMs examined the application of different sequential methods for change-point detection (i.e., Shiryaev, Shiryaev–Roberts, CUSUM, and M-in-a-row) as a means to detect learning. Subsequent developments in the area include various learning models with or without covariates, as well as different estimation algorithms and strategies. To be more practically viable, future research in this area should include more realistic but challenging scenarios such as measuring a large number of attributes at multiple time points, as well as a closer examination of how technology can be harnessed not only to measure, but also to facilitate learning.

To conclude, we concur with a recent review that, although CDM research to date has produced a large number of methodological advances, it has not yet fulfilled its promise of facilitating formative assessment in the school context. Fully developing the three lines of research discussed above, together with the current advances, can pave the way for CDM to have a real impact on everyday teaching and learning. A glimpse of this possibly can be found in a small quasi-experimental study, where an online tutoring program that generated individualized remedial learning materials in conjunction with a CDM diagnostic report was evaluated. The experimental group had a remedial class in the multimedia classroom using the tutoring program the week after the pretest, whereas the control group received the traditional group-based remedial instruction in their

classroom from teachers who were also given the same diagnostic report. When both groups were assessed again, students in the experimental group outperformed those in the control group, which corroborates the findings of a similar study. More importantly, the individualized instruction was more beneficial for medium- and low-achieving students – this finding has important equity implications for many students who do not have access to quality teachers. More empirical studies involving assessments specifically designed to be cognitively diagnostic, that highlight the practical benefits of CDM, need to be carried out. Not only will such studies spur further applications of CDM, they will also generate new lines of research as novel problems emerge from these applications.

9.6 Related Literature

In the first decade of this century, several reviews and books were published documenting the progress made in the area of CDM. These include DiBello, Roussos, and Stout (2006) and Rupp and Templin (2008), among others. Much of the content of these reviews is still relevant today. However, the purpose of this chapter was to summarize the more recent developments since their publication. For this purpose, the work by Sorrel *et al.* (2016), which documents the important steps in CDM applications from the Q-matrix development to gathering reliability and validity evidences, was taken as a reference point. There have also been several articles discussing the usefulness of CDM as a measurement tool for providing diagnostic feedback in education. The interested reader can refer to de la Torre (2012), de la Torre and Minchen (2014), Leighton and Gierl (2007), and Nichols, Chipman, and Brennan (1995), among others. In addition to education (e.g., Tjoe & de la Torre, 2014; Wu, 2019), as discussed in the chapter, other empirical applications have emerged in areas such as clinical psychology (e.g., de la Torre, van der Ark, & Rossi, 2018; Templin & Henson, 2006) or industrial-organization psychology (e.g., Bley, 2017; J. Chen & Zhou, 2017). These papers provide an overview of how CDMs are being applied in the real world. Much of the empirical work available was reviewed by Sessoms and Henson (2018), which is a good starting point to understand characteristics of practical applications of CDMs.

There is a large number of articles dedicated to the development of new models. The interested reader may refer to de la Torre (2009b) for a didactic introduction to the DINA model and its estimation using MML, where the G-DINA model is a natural extension (de la Torre, 2011). The G-DINA model is a general framework that subsumes several of the most popular CDMs available for dichotomous data and attributes. The framework has been extended to the case of polytomous attributes (J. Chen & de la Torre, 2013), polytomous data (de la Torre, 2009a; Ma & de la Torre, 2016; Ozaki, 2015), and continuous response (Minchen & de la Torre, 2018; Minchen, de la Torre, & Liu, 2017). Other general CDMs, such as the general diagnostic model (von Davier, 2005) and the log-linear cognitive diagnosis model (Henson, Templin, & Willse, 2009), also exist. Works related to hierarchical attribute structures can be found in Akbay and de la Torre (2020)

and Tu *et al.* (2019). For some of the latest developments in MML estimation in the CDM context, see Ma and Jiang (2021). A number of works have been published on model identifiability in recent years. Authors wishing to extend what has been discussed in this chapter can find more information in Chiu, Douglas, and Li (2009), Gu and Xu (2021), Xu (2017), among others. Finally, for an introduction to MCMC estimation, see de la Torre and Douglas (2004), Henson, Templin, and Willse (2009), and Liu and Johnson (2019). Readers interested in nonparametric approaches can refer to the original article by Chiu and Douglas (2013) and the generalization of the method in Chiu, Sun, and Bian (2018).

Accurate attribute classification hinges on the correct specification of the Q-matrix (de la Torre, Hong, & Deng, 2010; Nájera, Sorrel, & Abad, 2019). Discussions on theory-based Q-matrix development can be found in Li and Suen (2013), Sorrel *et al.* (2016), and Tjoe and de la Torre (2014); in contrast, the exploratory or empirically based Q-matrix development procedures are discussed in Y. Chen *et al.* (2015, 2018a), Culpepper (2019), and Liu, Xu, and Ying (2012). Researchers have also started looking into the determination of the number of attributes (Nájera, Abad, & Sorrel, 2021; Robitzsch & George, 2019; Xu & Shang, 2018) as test measures. Several procedures for improving provisional Q-matrices have been proposed, and these include the method based on the general discrimination index (de la Torre & Chiu, 2016), a sequential method using the Wald test (Ma & de la Torre, 2020a), and the more recent Hull method (Nájera *et al.*, 2020). Other methods, such as the nonparametric method (Chiu, 2013) or those based on model fit information (e.g., Chen, 2017; Kang, Yang, & Zeng, 2019), have also been developed.

With respect to the literature on model fit evaluation, the study by Chen, de la Torre, and Zhang (2013) explores several statistics for evaluating both absolute fit and relative fit. Readers can also refer to Hansen *et al.* (2016), Liu, Tian, and Xin (2016), and Sorrel *et al.* (2017a) for works related to absolute fit evaluation, and de la Torre and Lee (2013), Ma, Iaconangelo, and de la Torre (2016), and Sorrel *et al.* (2017b) for relative fit evaluation at the item and test levels.

Huebner and Wang (2011) discussed various approaches for classifying individuals. A growing literature focussing on assessing the reliability of these classifications has been growing lately. Initial ideas on this topic are discussed in Cui, Gierl, and Chang (2012), Templin and Bradshaw (2013), and W. Wang *et al.* (2015). More recent proposals on assessing reliability can be found in Iaconangelo (2017) and Sinharay and Johnson (2019), and a summary of the approaches in Johnson and Sinharay (2020). With respect to validity evidence as it pertains to test fairness, a number of works, which include the use of traditional fit statistics (e.g., Hou, de la Torre, & Nandakumar, 2014; Qiu, Li, & Wang, 2019), as well as Bayesian approaches (X. Li & Wang, 2015), have been published. In addition to routine analysis, the need for empirical studies to gather evidence to support the valid use of CDM scores has been emphasized (Sessoms & Henson, 2018). Examples of these studies that include validity evidence can be found in de la Torre, van der Ark, and Rossi (2018), Ren *et al.* (2021), and Sorrel *et al.* (2016).

With respect to CD-CAT, one of the first available works is that of Cheng (2009). Several studies have continued the development of CD-CAT methodologies – new item selection rules and optimal criteria (e.g., Kaplan, de la Torre, & Barrada, 2015; Xu, Wang, & Shang, 2016; H. D. Yigit, Sorrel, & de la Torre, 2019), stopping rules (e.g., Guo & Zheng, 2019; Hsu, Wang, & Chen, 2013), and control of exposure and content balancing procedures (e.g., C. Wang, Chang, & Douglas, 2012; C. Wang, Chang, & Huebner, 2011; C. Zheng & Wang, 2017) have been proposed. Other studies have explored aspects related to item bank calibration (e.g., Huang, 2018; Sorrel, Abad, & Nájera, 2021) or procedures for updating item banks (e.g., P. Chen *et al.*, 2012; Wang, Cai, & Tu, 2020). Regarding a related topic (i.e., optimal test assembly), the interested reader can consult the works of Finkelman *et al.* (2010), Finkelman, de la Torre, and Karp (2020), Kuo, Pai, and de la Torre (2016), and Lin, Gong, and Zhang (2017). The literature related to the measurement of learning in the context of CDM remains scant to date. Examples of works in this area include Y. Chen *et al.* (2018b), S. Wang *et al.* (2018), Ye *et al.* (2016), H. Yigit and Douglas (2021), and Zhang and Chang (2020).

Finally, the available references that specifically address the existing software packages will be discussed, most of which are in the form of R packages (R Core Team, 2013). Two general packages, namely, `GDINA` (Ma & de la Torre, 2020b) and `CDM` (George *et al.*, 2016), are specifically designed for CDM analyses. Other more specific packages have also been developed: `cdmTools` deals with dimensionality determination and Q-matrix specification (Nájera, Abad, & Sorrel, 2021); `dina` (Culpepper & Balamuta, 2015) and `rrum` (Culpepper, Hudson, & Balamuta, 2019), which are based on Culpepper (2015) and Culpepper and Hudson (2018), respectively, can be used for the Bayesian estimation of the DINA model and the RRUM. Classification based on nonparametric methods can be carried out using `NPCD` (Y. Zheng & Chiu, 2019). Readers wishing to estimate CDMs using software other than R can refer to Templin and Hoffman (2013) and Zhan *et al.* (2019) for tutorials on how CDM analysis can be implemented in Mplus (Muthén & Muthén, 1998–2017) and the JAGS program (Plummer, 2003), respectively. Several of these programs have been compared in terms of usability, analysis types, and output, among others (Sen & Terzi, 2020). Finally, the R package `cdcatR` presented in Sorrel, Abad, and Nájera (2021) can be used to perform CD-CAT studies.

References

Akbay, L., & de la Torre, J. (2020). Estimation approaches in cognitive diagnosis modeling when attributes are hierarchically structured. *Psicothema*, *32*(1), 122–129.

Bley, S. (2017). Developing and validating a technology-based diagnostic assessment using the evidence-centered game design approach: An example of intrapreneurship competence. *Empirical Research in Vocational Education and Training*, *9*(1), 6.

Chen, J. (2017). A residual-based approach to validate q-matrix specifications. *Applied Psychological Measurement*, *41*(4), 277–293.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*(6), 419–437.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123–140.

Chen, J., & Zhou, H. (2017). Test designs and modeling under the general nominal diagnosis model framework. *PloS one*, *12*(6), e0180016.

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, *77*(2), 201–222.

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018a). Bayesian estimation of the DINA Q matrix. *Psychometrika*, *83*(1), 89–108.

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2018b). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement*, *42*(1), 5–23.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: Cd-cat. *Psychometrika*, *74*(4), 619.

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618.

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*(2), 225–250.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633.

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*(2), 355–375.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19–38.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476.

Culpepper, S. A. (2019). Estimating the cognitive diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika*, *84*(2), 333–357.

Culpepper, S. A., & Balamuta, J. J. (2015). dina: Bayesian estimation of DINA model [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dina (R package version 2.0.0)

Culpepper, S. A., & Hudson, A. (2018). An improved strategy for Bayesian estimation of the reduced reparameterized unified model. *Applied Psychological Measurement*, *42*(2), 99–115.

Culpepper, S. A., Hudson, A., & Balamuta, J. J. (2019). rrum: Bayesian estimation of the reduced reparameterized unified model with Gibbs sampling [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rrum (R package version 0.2.0)

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J. (2012). Application of the DINA model framework to enhance assessment and learning. In M. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-Pacific* (pp. 92–110). New York: Springer.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical q-matrix validation. *Psychometrika*, *81*(2), 253–273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*(2), 227–249.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355–373.

de la Torre, J., & Minchen, N. D. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Educational Psychology*, *20*(2), 89–97.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281–296.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics*, *26*, 979–1030.

Finkelman, M. D., de la Torre, J., & Karp, J. A. (2020). Cognitive diagnosis models and automated test assembly: an approach incorporating response times. *International Journal of Testing*, *20*(4), 299–320.

Finkelman, M. D., Kim, W., Roussos, L., & Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Applied Psychological Measurement*, *34*(5), 310–326.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package cdm for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1–24.

Gu, Y., & Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica*, *31*(1), 449–472.

Guo, L., & Zheng, C. (2019). Termination rules for variable-length CD-CAT from the information theory perspective. *Frontiers in Psychology*, *10*, 1122.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate dif in the DINA model. *Journal of Educational Measurement*, *51*(1), 98–125.

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, *37*(7), 563–582.

Huang, H.-Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification*, *35*(3), 437–465.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*(2), 407–419.

Iaconangelo, C. J. (2017). *Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models* (Unpublished doctoral dissertation). Rutgers University-School of Graduate Studies.

Johnson, M. S., & Sinharay, S. (2020). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics*, *45*(1), 5–31.

Kang, C., Yang, Y., & Zeng, P. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement*, *43*(7), 527–542.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *39*(3), 167–188.

Kuo, B.-C., Pai, H.-S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*(5), 315–330.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.

Li, H., & Suen, H. K. (2013). Constructing and validating a q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1–25.

Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28–54.

Lin, Y., Gong, Y.-J., & Zhang, J. (2017). An adaptive ant colony optimization algorithm for constructing cognitive diagnosis tests. *Applied Soft Computing*, *52*, 1–13.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*(7), 548–546.

Liu, X., & Johnson, M. S. (2019). Estimating CDMS using MCMC. In *Handbook of diagnostic classification models* (pp. 629–646). New York: Springer.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*(1), 3–26.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 253–275.

Ma, W., & de la Torre, J. (2020a). An empirical q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 142–163.

Ma, W., & de la Torre, J. (2020b). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*(3), 200–217.

Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied Psychological Measurement*, *45*(2), 95–111.

Minchen, N. D., & de la Torre, J. (2018). A general cognitive diagnosis model for continuous-response data. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 30–44.

Minchen, N. D., de la Torre, J., & Liu, Y. (2017). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics*, *42*(6), 651–677.

Muthén, L., & Muthén, B. (1998-2017). *Mplus user's guide*, 8th ed. Los Angeles, CA: Muthén & Muthén.

Nájera, P., Abad, F. J., & Sorrel, M. A. (2021). Determining the number of attributes in cognitive diagnosis modeling. *Frontiers in Psychology*, *12*, 321.

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical q-matrix validation. *Educational and Psychological Measurement*, *79*(4), 727–753.

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2020). Balancing fit and parsimony to improve q-matrix validation. *British Journal of Mathematical and Statistical Psychology*, *74*(S1), 110–130.

Nichols, P., Chipman, S., & Brennan, R. (1995). Cognitive structure testing: A computer system for diagnosis of expert-novice differences. In Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.)*, Cognitively Diagnostic Assessment* (pp. 251–278). Abingdon: Routledge.

Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement*, *39*(6), 431–447.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using GIBBS sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, pp. 1–10).

Qiu, X.-L., Li, X., & Wang, W.-C. (2019). Differential item functioning in diagnostic classification models. In *Handbook of Diagnostic Classification Models* (pp. 379–393). Berlin: Springer.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from www.R-project.org/

Ren, H., Xu, N., Lin, Y., Zhang, S., & Yang, T. (2021). Remedial teaching and learning from a cognitive diagnostic model perspective: Taking the data distribution characteristics as an example. *Frontiers in Psychology*, *12*.

Robitzsch, A., & George, A. C. (2019). The R package CDM for diagnostic modeling. In *Handbook of Diagnostic Classification Models* (pp. 549–572). Berlin: Springer.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219–262.

Sen, S., & Terzi, R. (2020). A comparison of software packages available for DINA model estimation. *Applied Psychological Measurement*, *44*(2), 150–164.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17.

Sinharay, S., & Johnson, M. S. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In *Handbook of Diagnostic Classification Models* (pp. 359–377). Berlin: Springer.

Sorrel, M. A., Abad, F. J., & Nájera, P. (2021). Improving accuracy and usage by correctly selecting: The effects of model selection in cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *45*(2), 112–129.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017a). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*(8), 614–631.

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017b). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *13*(S1), 39.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*(3), 506–532.

Templin, J. L., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287.

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50.

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, *26*(2), 237–255.

Tu, D., Wang, S., Cai, Y., Douglas, J., & Chang, H.-H. (2019). Cognitive diagnostic models with attribute hierarchies: Model estimation with a restricted q-matrix design. *Applied Psychological Measurement*, *43*(4), 255–271.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2).

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*(1), 95–109.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*(3), 255–273.

Wang, D., Cai, Y., & Tu, D. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known q-matrix. *Multivariate Behavioral Research*, 1–13.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 57–87.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457–476.

Wu, H.-M. (2019). Online individualised tutor for improving mathematics learning: A cognitive diagnostic model approach. *Educational Psychology*, *39*(10), 1218–1232.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, *45*(2), 675–707.

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*(523), 1284–1295.

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 291–315.

Ye, S., Fellouris, G., Culpepper, S. A., & Douglas, J. (2016). Sequential detection of learning in cognitive diagnosis. *British Journal of Mathematical and Statistical Psychology*, *69*(2), 139–158.

Yigit, H., & Douglas, J. (2021). First-order learning models with the GDINA: Estimation with the EM algorithm and applications. *Applied Psychological Measurement*, doi.org/10.1177/0146621621990746.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement*, *43*(5), 388–401.

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*, *44*(4), 473–503.

Zhang, S., & Chang, H.-H. (2020). A multilevel logistic hidden Markov model for learning under cognitive diagnosis. *Behavior Research Methods*, *52*(1), 408–421.

Zheng, C., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *41*(7), 561–576.

Zheng, Y., & Chiu, C.-Y. (2019). Npcd: Nonparametric methods for cognitive diagnosis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=NPCD (R package version 1.0-11)

10 Encoding Models in Neuroimaging

Fabián A. Soto and F. Gregory Ashby

10.1 Introduction

One of the greatest barriers to progress in mathematical psychology is model mimicry. In almost every domain of cognitive modeling, there are competing models that assume qualitatively different perceptual and cognitive processes, yet are able to mimic the behavioral predictions of each other. One reason for this is that although competing models may make very detailed predictions about psychological processes, historically those processes have been unobservable and, as a result, the models are tested only against crude dependent measures, such as response accuracy and response time.

Within the past few decades, a wide variety of new neuroimaging technologies have been developed that allow levels of observability into human brain function that seemed unimaginable when many currently popular mathematical models in psychology were first proposed. Included in this list are functional magnetic

resonance imaging (fMRI), positron emission tomagraphy (PET), magnetoencephalography (MEG), functional near-infrared spectroscopy (fNIRS), electrocorticography (ECoG), and high-resolution electroencephalography (EEG). Although these methods all have limitations, they nevertheless have the potential to allow unprecedented observability into the perceptual and cognitive processes predicted to underlie competing mathematical models of perception and cognition. As a result, testing models against neuroimaging data in addition to the more traditional response accuracies and response times offers an exciting possible solution to the model mimicry problems that plague mathematical psychology.

Despite their promise, neuroimaging data are infrequently used to test mathematical models of the type that are common in mathematical psychology. There are several reasons for this. First, neuroimaging is still a relatively new technology and neuroimaging data analysis is still in a period of rapid development. Second, all of these neuroimaging technologies were developed outside of mathematical psychology. Third, most models in mathematical psychology make few, if any, neuroscience predictions. At first glance, the latter of these reasons seems the most limiting, but in fact, several data analysis methods that were developed to analyze fMRI data can be used to test models that make no neuroscience assumptions. Included in this list are *model-based fMRI* and *representational similarity analysis* (RSA).

All neuroimaging technologies work in a similar way. In all cases, recordings are collected at discrete times and locations in the brain while the subject is engaged in some perceptual or cognitive task. The recordings are directly (e.g., ECoG, EEG) or indirectly (e.g., fMRI, PET) related to neural activation. The spatial resolution varies. ECoG can sometimes measure action potentials in single neurons, whereas each EEG electrode is influenced by millions of neurons. Temporal resolution also varies, with ECoG, EEG, and MEG at one extreme (with resolutions near 1 ms) and PET at the other (with resolutions of 5–10 s). State-of-the-art functional MRI scanners, with multi-band slice acquisition, have a temporal resolution of about 500 ms and a spatial resolution of 1–2 mm (i.e., limited by the point-spread function of the blood oxygen level-dependent, or BOLD response; Fracasso, Dumoulin, & Petridou, 2021).

In general, neuroimaging data analysis techniques can be classified as either *encoding* or *decoding* methods. Encoding methods use knowledge of the experimental design and stimuli to build a model that predicts the neural activation that should be generated at each recording site on every trial. Decoding methods refer to approaches that make inferences in the opposite direction – that is, they use the observed recordings to make predictions about stimuli and other events in the experiment (Haynes & Rees, 2006; Naselaris *et al.*, 2011; Norman *et al.*, 2006; Pereira, Mitchell, & Botvinick, 2009). The idea is that if a brain region of interest (ROI) responds differently to two different stimulus attributes then that ROI might be processing those attributes differently. The most widely used decoding method is known as pattern classification or even more commonly as multivoxel pattern analysis (MVPA).

Encoding models are similar to traditional models in mathematical psychology. To model behavior in a task, a mathematical psychologist will typically combine assumptions about the underlying perceptual and cognitive processes with knowledge of the task to write equations that predict the participant's accuracy and/or response time. To build an encoding model, assumptions about the underlying neural processes are combined with knowledge of the task and the type of neuroimaging technique being used to write equations that predict values of the dependent variable that is measured at each recording site. For example, an encoding model of fMRI data would predict the observed BOLD response at each voxel in response to each stimulus presentation. Forward inferences of this type are used for two primary purposes. First, they can be used to identify brain regions that are sensitive to specific attributes of the stimulus events. For example, when natural scenes are described by the outputs of many phase-invariant Gabor filters, simple fMRI encoding models accurately predict the BOLD response in early visual areas, but not in high-level areas of the visual cortex (Kay *et al.*, 2008; Naselaris *et al.*, 2009). In contrast, when the same scenes are described using semantic category labels, encoding models accurately predict activation in high-level visual areas but not in the early visual cortex (Mitchell *et al.*, 2008; Naselaris *et al.*, 2009). Second, encoding models can be used to test theories of cognitive and neural processing. If a theory accurately describes the cognitive and neural processing that occurs during a specific task, then it should be possible to use that theory to construct an encoding model that accurately predicts the dependent variables recorded in a set of pre-specified ROIs.

Because these two goals are somewhat different, it is not surprising that a diverse set of encoding models have been proposed (e.g., Ashby, 2019). The most widely used fMRI encoding model is the familiar general linear model (GLM) from statistics, which is used most commonly to identify brain regions that are sensitive to the simplest possible attribute of a stimulus event – namely, its presence or absence. All other encoding models are more ambitious. Arguably the next most popular fMRI encoding approach is dynamic causal modeling (DCM), which identifies a candidate set of brain regions that mediate event processing, along with all of their functional interconnections (Ashby, 2019; Friston, Harrison, & Penny, 2003). DCM is also more complex than other encoding models, partly because it uses a nonlinear model relating the BOLD response to neural activation and partly because it uses a variational Bayesian approach for model selection.

The vast majority of encoding models were developed to be tested against fMRI data. Even so, for the most part, the models can all be applied to any neuroimaging technology. The only significant difference from one technology to another is in the interface that converts predicted activation in a neural population to values of the dependent variable that the technology measures. For example, in the case of fMRI data, one needs to model the transformation from neural activation to the BOLD response recorded in fMRI experiments. With EEG data, one needs to include a head model that accounts for electromagnetic properties of the head and of the sensor array. But in all cases, the model of each neural population and of how

the population activations are combined is roughly the same. However, because the models we discuss were developed for application to fMRI data, we will assume an fMRI application in the rest of this chapter. For most of the chapter, this just means that we will refer to a recording site as a voxel, and the time between recordings as the TR (repetition time; the amount of time it takes the scanner to measure BOLD responses from all voxels in the brain). Except for this nomenclature, the only part of the chapter unique to fMRI is discussed in the subsection entitled "Linking Neural Activation to the BOLD Response," which considers the interface between the neural activations predicted by the models and the dependent variable most commonly measured in fMRI experiments.

10.2 Voxel-Based Encoding Models

Encoding models fall into two general classes: those that were constructed specifically to analyze fMRI data, and models that were originally designed for other purposes. The former class are often called *voxel-based encoding models*. The latter class can take many forms – from purely cognitive models of the type that are common in mathematical psychology to models with considerable biological detail (a branch of modeling called computational cognitive neuroscience; e.g., Ashby, 2018). FMRI data are used along with a variety of other data types to test and refine these models. The process of testing such models against fMRI data is known as model-based fMRI. We consider model-based fMRI later in the chapter. This section describes voxel-based encoding models.

Voxel-based encoding models encompass a variety of different models, but they all share enough features to be characterized within a single framework. As we will see in this section, all current voxel-based encoding models include an encoding model that predicts how every hypothesized neural population responds to each stimulus, and a measurement model that first transforms neural population responses into aggregate neural activity and then into values of the dependent variable being measured (e.g., the fMRI BOLD response). While most encoding models include a highly nonlinear transformation from stimulus to neural response, the measurement model is usually linear, and such models are often referred to as linearized encoding models. This means that most voxel-based encoding models can be seen as instances of linear regression with basis functions (Hastie, Tibshirani, & Friedman, 2009).

10.2.1 Encoding Model

Encoding models begin with a mathematical description of the relation between a set of stimuli S_i, with $i = 1, 2, \ldots, N_s$, and the response of a neural channel r_c, with $c = 1, 2, \ldots, N_c$. Neural channels can represent either a single neuron or a population of neurons with similar properties, with the latter option being more common in the computational neuroimaging literature. Most encoding models

assume that the channel response depends on the identity of the stimulus S_i, certain channel tuning parameters, various state variables, and properties of the neural noise. The tuning parameters, which are collected in the vector $\underline{\boldsymbol{\theta}}$, include, for example, constants that determine the channel's maximum possible response, and its preferred stimulus. The state variables, collected in the vector $\underline{\mathbf{x}}$, include other variables that could affect the channel response, including, for example, the responses of other channels in the population. Given these definitions, the standard approach is to first define the mean channel response

$$\mathrm{E}\left[r_c|\mathrm{S}_i\right] = f_c\left(\mathrm{S}_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right), \tag{10.1}$$

where E denotes expected value, and f_c is the channel tuning function, which is specified as part of the model. Tuning functions are discussed in more detail below, but it is important to note that in many applications, the alternative encoding models that are tested against data are identical, except for their tuning functions.

Most encoding models assume that channels operate in the presence of noise, but they differ in how that noise is modeled. One approach is to assume that the response of channel c to presentation of stimulus S_i is

$$r_c(\mathrm{S}_i) = f_c\left(\mathrm{S}_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right) + \epsilon_c, \tag{10.2}$$

where ϵ_c is zero-mean noise (e.g., Pouget, Dayan, & Zemel, 2000). A common choice is to assume Gaussian noise with some fixed variance. Note that this model predicts that the variance of the channel response does not change as the mean response increases. There is support for this assumption in channels that include a large population of neurons (Y. Chen, Geisler, & Seidemann, 2006), but in single neurons, the variance of the spike count tends to increase in proportion to the mean (e.g., Tolhurst, Movshon, & Dean, 1983). Therefore, the fixed-variance Gaussian model is most appropriate when modeling channels of many neurons. A popular approach to modeling channels in which the variance of the response increases with the mean is to assume that r_c is Poisson distributed with mean $f_c\left(\mathrm{S}_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right)$ (e.g., Zemel, Dayan, & Pouget, 1998). Therefore, this model assumes that the channel response has probability density function

$$P[r_c|\mathrm{S}_i] = \frac{f_c\left(\mathrm{S}_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right)^{r_c} \mathrm{e}^{-f_c\left(\mathrm{S}_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right)}}{r_c!}. \tag{10.3}$$

Because the variance of a Poisson distribution equals its mean, this model predicts that the variance of the channel response increases with the mean response. Note that Equations (10.2) and (10.3) both assume that the mean channel response satisfies Equation (10.1).

Most models include multiple channels, each described by a version of Equations (10.1) and (10.2) or Equations (10.1) and (10.3), and which are combined into a random vector of responses $\underline{\mathbf{r}} = \left[r_1, r_2, \ldots, r_{N_c}\right]$ that describe the response of all N_c channels to the presented stimulus. This is known as a population encoding model (Pouget, Dayan, & Zemel, 2000, 2003), and $\underline{\mathbf{r}}$ is usually referred to as a *population response*. In particular, voxel-based encoding models assume that

every voxel includes a mixture of various populations of neurons, and that each population is tuned to a different attribute of the stimulus. The populations are commonly referred to as channels. For example, the most primitive visual encoding model might assume that each population or channel is tuned to a Gabor patch of a certain spatial frequency and orientation. But the populations could be tuned to anything. At the opposite extreme, they might be tuned to semantic category labels, such as rock, ocean, table, chair, or lamp. Voxel-based encoding models are most commonly used to identify brain regions that are sensitive to these attributes, so it is not unusual to build multiple encoding models for the same data that are each sensitive to a different set of stimulus attributes.

We can make this more concrete with an example of what has been termed the standard model of dimension encoding (Pouget, Dayan, & Zemel, 2000, 2003). This model is typically restricted to applications in which the stimuli vary on a single dimension. Suppose the numerical value of stimulus S_i on this dimension is s_i. The model assumes Gaussian tuning functions, so in this case it predicts that

$$f_c\left(s_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right) = r_c^{max} \exp\left[-\frac{1}{2}\left(\frac{s_i - s_c}{\omega_c}\right)^2\right], \tag{10.4}$$

where r_c^{max} represents the maximum response for channel c, s_c represents the value of the channel's preferred stimulus (i.e., the value of the stimulus that produces the channel's largest response), and ω_c represents the width of the tuning function. Many applications assume that all tuning functions have the same width (i.e., $\omega_c = \omega$, for all c), which is known as the homogeneous standard model. In all versions of the model, however, the channel tuning parameters are gathered together in the vector $\underline{\boldsymbol{\theta}}_c^{\top} = \left[r_c^{max}, s_c, \omega_c\right]^{\top}$, where $\top$ denotes transpose. Note that in this case, the state vector $\underline{\mathbf{x}}$ is empty. Also note that this model makes it possible to predict the mean channel responses as soon as the stimuli are selected, and therefore, before data collection begins.

Figure 10.1a shows the tuning functions of a large collection of channels from a typical application of this standard one-dimensional model. Note that all channels have identical shape ($r_c^{max} = r^{max}$ and $\omega_c = \omega$) and that the preferred stimuli for the various channels are evenly spaced on the stimulus dimension ($s_c = s_{c-1} + k$, for some small constant k). The shape of the tuning functions for all channels is therefore characterized by a single canonical tuning curve.

Now imagine presenting a specific stimulus S_i to the model and recording the response of all N_c channels in the population response vector $\underline{\mathbf{r}}$. A convenient way to describe these responses is via a *population response plot*, in which neural responses are plotted on the ordinate and the numerical values of each channel's preferred stimulus are plotted on the abscissa. Figure 10.1b shows the population response of the model in Figure 10.1a to a stimulus with value 0. Each solid dot shows the response of a different channel in the absence of noise, and each open dot denotes a possible response of the same channels in the presence of noise.

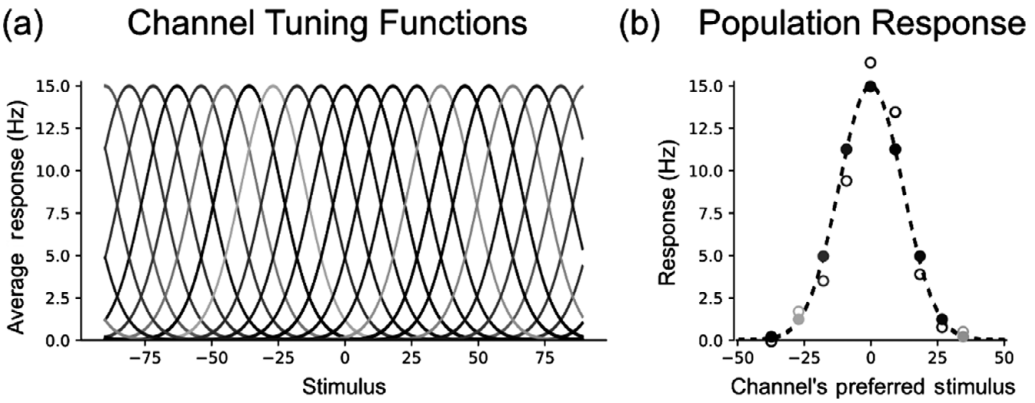


Figure 10.1 *The standard model of dimension encoding. Panel (a) shows the tuning curves of the various channels included in the model. The peak of each tuning curve is centered at the channel's preferred stimulus value. Panel (b) shows the population response plot of this model on a hypothetical trial when a stimulus with value 0 is presented. Each solid dot shows the response of a different channel in the absence of noise, and each open dot denotes a possible response of the same channels in the presence of noise.*

Note that, because all channels have the same width and are equally spaced on the stimulus dimension, the expected population response has the same shape as the canonical tuning function. This property of the standard encoding model is a continuous source of confusion for both experimentalists and modelers, who sometimes confuse population response plots with tuning functions in their interpretation of encoding models. A population response function with the same shape as the canonical tuning function is not a general property of encoding models, but arises specifically from the homogeneous model (i.e., in which all tuning functions are identical, except for their preferred stimulus).

Channel noise distributions have been estimated empirically, and there is evidence that humans use knowledge of this uncertainty during perceptual decision-making (Van Bergen *et al.*, 2015). Even so, it is common in the cognitive neuroscience literature to find applications in which channel noise is not modeled, with responses being described simply by Equation (10.1). Within the general framework presented here, those applications implicitly assume Equation (10.2) and Gaussian noise with a variance that is invariant across channels. When channel noise is modeled, a common assumption is that the noise is independently and identically distributed across multiple channels. In contrast, as mentioned earlier, some approaches model the channel response as Poisson distributed [i.e., as in Equation (10.3)], which scales the noise variance up with the mean channel response.

Of course, there are a variety of ways to construct more complex models. First, the model is easily extended to multidimensional stimuli. For example, in vision research it is common to represent images as two-dimensional matrices

of pixel values, with each channel's tuning function being defined in that space. Many models represent the operation of primary visual cortex, or V1, through a large population of channels in which the tuning function of each channel is a Gabor wavelet tuned to a certain specific spatial location, orientation, and spatial frequency (e.g., Kay *et al.*, 2008; Naselaris *et al.*, 2009). In their structural encoding model, Naselaris *et al.* (2009; see also Kay *et al.*, 2008) assumed a total of 10,921 such channels.

The Gabor wavelet model of tuning functions is based on years of research on the response properties of neurons in V1. The tuning properties of channels in higher visual areas are less well understood. As a result, in applications that depend on a participant's perceptual or cognitive impressions of a set of images, a more generic tuning function might be more appropriate. The Gaussian tuning function of Equation (10.4) is easily generalized to any arbitrary multidimensional stimuli. For example, consider a set of stimuli that vary on multiple dimensions and a channel in which the preferred stimulus is S_c. Then a multidimensional analog of Equation (10.4) assumes that the channel response to stimulus S_i is

$$f_c\left(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right) = r_c^{max} \exp\left[-\frac{1}{2}\left(\frac{\Delta(S_i, S_c)}{\omega_c}\right)^2\right], \tag{10.5}$$

where $\Delta(S_i, S_c)$ is the distance in perceptual space between the representations of stimuli S_i and S_c. Equation (10.5), which is an example of a radial basis function (e.g., Buhmann, 2003), is a popular method for modeling the receptive fields of sensory units in many different modeling approaches (e.g., Ashby, Ennis, & Spiering, 2007; Kruschke, 1992).

A second approach to building a more complex model is to express channel tuning via a composite function: $f_c\left(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right) = g_{c2}\left[g_{c1}\left(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right)\right]$. For example, in the Naselaris *et al.* (2009) model, the channel response is determined by applying a compressive nonlinearity to the output of the Gabor wavelet. If we denote the response of Gabor wavelet c to image S_i as $g_c(S_i)$, then according to this model the response of channel c is

$$f_c\left(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right) = \log[g_c(S_i) + 1]. \tag{10.6}$$

The +1 just ensures that the channel response is never negative. Because log is a negatively accelerating function, this transformation models response compression at the neural level.

A third common generalization of the standard model is to assume that the channel response depends on state variables indexed in the vector $\underline{\mathbf{x}}$. For example, $\underline{\mathbf{x}}$ might include the responses of other channels in the population. In this case, a popular approach is to use these other responses to normalize the response of each channel:

$$f_c\left(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right) = \frac{g_c\left(S_i\right)^\nu}{\kappa^\nu + \sqrt{\sum_j \alpha_j [g_j\left(S_i\right)]^\nu}}. \tag{10.7}$$

This is called *divisive normalization*, and it is a ubiquitous computation in cortical circuits (Carandini & Heeger, 2012). In this model, the channel response is normalized by a weighted sum of the response of all channels. The weights α_j represent the level to which other channels suppress the response of channel c, ν increases competition among channels for activation, and κ prevents division by zero.

10.2.2 Measurement Model

The encoding models discussed so far describe activity in each channel. However, in most applications, the individual channel responses are assumed to be unobservable. For example, in applications to fMRI, the BOLD response recorded in each voxel is assumed to be a mixture of many channel responses. Therefore, to test encoding models against empirical data, a model interface is required that specifies how the channels combine to determine the value of the dependent variable of interest (see Van Bergen *et al.*, 2015). This interface is called the *measurement model*.

The measurement model must solve two separate problems. First, even with state-of-the-art high-resolution MRI scanners, each voxel includes many neurons, and therefore presumably many different neural channels. Therefore, the first problem is to model how the various hypothesized channels combine to determine the amplitude of the neural activation that drives the BOLD response in each voxel.

Second, in the encoding models considered so far, the channel response $r_c(\mathrm{S}_i)$ is a single value that is presumed to represent the amplitude of neural activation in channel c when stimulus S_i is presented. In contrast, the BOLD response recorded from each voxel when stimulus S_i is presented is a time series that persists for 30 s or so and depends in a complicated way on concentrations of oxygenated and deoxygenated hemoglobin, cerebral blood flow, and venous blood volume (Buxton, 2013). Neural activation increases the BOLD response, but the BOLD response is only an indirect measure of neural activation (Ogawa *et al.*, 1990a, 1990b). So the second problem in applications of encoding models to fMRI data is to link the neural activation values predicted by the models to the observed BOLD time series recorded in fMRI experiments.

This section considers each of these problems in turn.

Aggregating Channel Responses

Each voxel in an fMRI experiment will include several hundred thousand neurons. As a result, any voxel-based encoding model that includes multiple channels will assume that every voxel in the ROI could potentially contain all of the hypothesized channels. This is true no matter how the channels are defined, although most models assume that the number of channels, and the number of neurons within each channel, are unknown. The most popular assumption is that the neural activation produced in a task-sensitive voxel in response to a stimulus presentation is a weighted linear combination of all the channels represented in that voxel,

where the weights are presumed to reflect the number of neurons within the voxel that contribute to each channel. Models in this class are often referred to as linearized encoding models because the measurement model assumes that the voxel-level neural activation is a weighted linear combination of the individual channel responses. When combined with a linear model of the relationship between neural activation and the observed BOLD response, such models can use the GLM for parameter estimation – that is, to estimate the values of the unknown weights that allow the model to give the best fits to the observed BOLD responses collected from that voxel on all TRs.

We can formalize these ideas as follows. Let $a_k(\mathrm{S}_i)$ denote the aggregate neural activity in voxel k to presentation of stimulus S_i, and let w_{ck} denote the contribution of channel c to this activity. Then the voxel-based (or linearized) encoding model assumes that

$$a_k(\mathrm{S}_i) = w_{1k} + \sum_{j=2}^{N_c} w_{jk}\, r_j(\mathrm{S}_i) + \epsilon_{m,k}, \tag{10.8}$$

where w_{1k} is the response of one channel in voxel k that gives the same constant response to all stimuli (to account for baseline activation that might occur in a voxel containing none of the hypothesized channels), and $\epsilon_{m,k}$ is the measurement error on channel k. The most common assumption is that $\epsilon_{m,k}$ is normally distributed with mean 0 and variance σ_m^2. This is called a linearized encoding model because it makes the simplifying assumption of a linear relation between channel responses and voxel activity. Note that this model predicts that the voxel activity $a_k(\mathrm{S}_i)$ is normally distributed or approximately normally distributed (in the Poisson case) with mean

$$\mathrm{E}[a_k(\mathrm{S}_i)] = w_{1k} + \sum_{j=2}^{N_c} w_{jk} f_c(\mathrm{S}_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}) \tag{10.9}$$

and in the case where the channels are independent, with variance

$$\mathrm{Var}[a_k(\mathrm{S}_i)] = \sigma_m^2 + \sum_{j=2}^{N_c} w_{jk}^2 \mathrm{Var}[r_j(\mathrm{S}_i)], \tag{10.10}$$

where $\mathrm{Var}[r_j(\mathrm{S}_i)]$ either equals σ_c^2 in the case of the Equation (10.2) Gaussian model or $f_c(\mathrm{S}_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}})$ in the case of the Equation (10.3) Poisson model.

Note that this model accounts for the separate contributions of the channel noise and the measurement noise [$\epsilon_{m,k}$ in Equation (10.8)] to the variability in $a_k(\mathrm{S}_i)$. In almost all cases, however, these will not be separately estimable. In fact, in linear models, it is well known that they are nonidentifiable. Instead, only the sum of these separate variances can be estimated (e.g., Ashby, 1992). As a result, in most applications, a single noise variance will be estimated and the source of the noise will be impossible to identify. Nevertheless, we include both noise sources for completeness.

Of course, there is a separate equation like Equation (10.8) for every stimulus in the ensemble. In all of these, the weights are identical because the weights are presumed to reflect the dominance of each channel within the voxel, which does not depend on what stimulus is presented. In contrast, the channel responses reflect the dominance of each feature within the stimulus, so these will change when the stimulus changes, but should be the same in all voxels. The standard way to keep track of all this is in matrix form. For example, consider an experiment with N_s different stimuli or events. The first step is to collect all channel responses – one for every channel – in an $N_s \times N_c$ channel-response matrix $\mathbf{R}$ defined as

$$\mathbf{R} = \begin{bmatrix} \underline{\mathbf{r}}(\mathrm{S}_1)^\top \\ \underline{\mathbf{r}}(\mathrm{S}_2)^\top \\ \vdots \\ \underline{\mathbf{r}}(\mathrm{S}_{N_s})^\top \end{bmatrix}. \tag{10.11}$$

So row i of $\mathbf{R}$ lists the population response to presentation of stimulus S_i, and column c lists the response of channel c to the presentation of each stimulus. If channel noise is modeled, then $\mathbf{R}$ is a random matrix. In most linearized encoding models, however, channel noise is not included and thus each channel is characterized by its mean response, computed as in Equation (10.1).

Encoding models assume that the channels and their tuning functions are known, so the mean channel response matrix $\mathrm{E}[\mathbf{R}]$ can be computed as soon as the stimulus set is selected, and therefore before the experiment begins. Voxel-based encoding models are therefore not used to estimate channel responses, because these are assumed to be known beforehand. Applying a voxel-based encoding model to neuroimaging data instead answers three different questions. First, it can identify the ROIs where the voxel activity most closely resembles the responses predicted by the set of presumed channels. Second, it provides an estimate of the relative frequency of each channel within every voxel. And third, for any single ROI it can tell whether the observed voxel activities are more consistent with one set of presumed channels or with another set.

The channel-response matrix described in Equation (10.11) accounts for the channel responses. The full set of model predictions can then be written in matrix form as

$$\begin{bmatrix} a_k(\mathrm{S}_1) \\ a_k(\mathrm{S}_2) \\ \vdots \\ a_k(\mathrm{S}_{N_s}) \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{r}}(\mathrm{S}_1)^\top \\ \underline{\mathbf{r}}(\mathrm{S}_2)^\top \\ \vdots \\ \underline{\mathbf{r}}(\mathrm{S}_{N_s})^\top \end{bmatrix} \begin{bmatrix} w_{1k} \\ w_{2k} \\ \vdots \\ w_{N_c k} \end{bmatrix} + \begin{bmatrix} \epsilon_{m,1} \\ \epsilon_{m,2} \\ \vdots \\ \epsilon_{m,N_s} \end{bmatrix},$$

or in shorthand form as

$$\underline{\mathbf{a}}_k = \mathbf{R}\underline{\mathbf{w}}_k + \underline{\boldsymbol{\epsilon}}_m, \tag{10.12}$$

where the random vector $\underline{\boldsymbol{\epsilon}}_m$ has a multivariate normal distribution with mean vector $\underline{\mathbf{0}}$ and variance–covariance matrix $\boldsymbol{\Sigma}_m$. Most applications assume that

$\mathbf{\Sigma}_m = \sigma_m^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and they also ignore channel noise, in which case $\mathbf{R}$ is replaced by E[$\mathbf{R}$]. In these cases, the only free parameters in the model are the weights $w_{1k}, w_{2k}, \ldots, w_{N_c k}$ and σ_m^2. Note that under these conditions, Equation (10.12) has exactly the same form as the GLM in statistics, which is usually stated as $\underline{\mathbf{y}} = \mathbf{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\epsilon}}$. As a result, if we assume that $\underline{\mathbf{a}}_k$ is linearly related to the observed BOLD response, then we can estimate the unknown weights in $\underline{\mathbf{w}}_k$ by solving the normal equations of the GLM (more on this shortly).

Equation (10.12) applies the encoding model to activity values from a single voxel. It is straightforward to extend the model to multiple voxels in an ROI. Adding more voxels does not change E[$\mathbf{R}$] since all voxels are exposed to the same stimulus events on every TR. Even so, the model allows two voxels to respond differently to the same stimulus because the channels might have different relative frequencies in the two voxels. So for every new voxel that is added, a new set of weights must be estimated. Mathematically, this is easily done by replacing the vector of weights $\mathbf{w}$ with a matrix $\mathbf{W}$ in which column k contains the weights associated with voxel k. The vector of voxel activities $\underline{\mathbf{a}}_k$ is expanded to a matrix $\mathbf{A}$ in which column k contains $\underline{\mathbf{a}}_k$ and we also need to add a new noise vector for each new voxel. These changes lead to the multivariate encoding model

$$\left[\underline{\mathbf{a}}_1 \quad \underline{\mathbf{a}}_2 \quad \cdots \quad \underline{\mathbf{a}}_{N_v}\right] = \mathbf{R}\left[\underline{\mathbf{w}}_1 \quad \underline{\mathbf{w}}_2 \quad \cdots \quad \underline{\mathbf{w}}_{N_v}\right] + \left[\underline{\boldsymbol{\epsilon}}_{m,1} \quad \underline{\boldsymbol{\epsilon}}_{m,2} \quad \cdots \quad \underline{\boldsymbol{\epsilon}}_{m,N_v}\right],$$

or in shorthand form

$$\mathbf{A} = \mathbf{RW} + \mathbf{E}_m. \tag{10.13}$$

When channel noise is ignored, this model is identical to the multivariate GLM. While each column of $\mathbf{A}$ represents a different activity profile (i.e., the vector of activities of a single voxel across stimulus conditions), each row of $\mathbf{A}$ represents a different *activity pattern,* or the vector of activities across multiple voxels in response to a single stimulus condition (Diedrichsen & Kriegeskorte, 2017). The distinction between activity profile and activity pattern at the level of voxels is analogous to the distinction between tuning function and population response at the level of neural channels.

Linking Neural Activation to the BOLD Response

As mentioned previously, the BOLD response is a time series. Active brain areas consume more oxygen than inactive areas, so when neural activity increases in an area, metabolic demands rise, and, as a result, oxygenated hemoglobin rushes into the area. Neural activity causes an immediate oxygen debt, and the resulting rush of oxygenated hemoglobin into the area causes the BOLD signal to rise quickly until it eventually reaches a peak at around 6 s after the neural activity that elicited these responses. After this peak, the BOLD signal gradually decays back to baseline over a period of 20–25 s (with the decay typically including a brief dip below baseline).

In contrast, the encoding models considered so far are static, in the sense that the predicted aggregate neural activity $a_k(\mathrm{S}_i)$ to presentation of stimulus S_i is a single value. All static encoding models make the same simplifying assumption that the

amplitude of the BOLD response in a voxel is proportional to the aggregated neural activation that occurs in that voxel. This enormously simplifies the problem of linking the aggregate activity predicted by the model to the observed BOLD response recorded in the experiment. The only remaining problem is to estimate a single amplitude of response from the BOLD time series. Furthermore, in most experiments, each stimulus will be presented multiple times, so there will be more than one such time series for stimulus S_i. Therefore, to apply a static encoding model, a single value that represents the amplitude of the BOLD response to stimulus S_i in voxel k must be estimated from these data. This problem is known in the neuroimaging literature as deconvolution or unmixing, and a solution to it is also required in decoding methods, such as multivoxel pattern analysis (MVPA). Not surprisingly, many alternative estimators have been proposed (e.g., Mumford *et al.*, 2012; Pedregosa *et al.*, 2015; B. O. Turner *et al.*, 2012).

In rapid event-related designs, which are the norm in modern fMRI research, stimuli are presented within 5 s or so of each other, as they are in most laboratory experiments. Since the BOLD response to neural activity might persist for 30 s, this means that the BOLD signals elicited by successive stimulus presentations will overlap in time. This overlap complicates the unmixing process. Mumford *et al.* (2012) proposed an efficient solution to this problem that they called least squares – separate (LSS). If there are N_E separate stimulus presentations, then LSS reruns the standard GLM regression analysis N_E separate times on the data from each voxel. In the ith of these N_E runs, the GLM includes two parameters – one regressor for the single trial on which the ith stimulus was presented and a second nuisance regressor that models the response to all other stimuli. The regression weight associated with the ith stimulus in this analysis is used as an estimate of the amplitude of the BOLD response in voxel k to the presentation of stimulus S_i. We will denote the BOLD time series in voxel k as $b_k(t)$ and the amplitude of this time series on trials when stimulus S_i is presented as $\tilde{b}_k(S_i)$. This LSS method was the most effective of a variety of alternative estimation methods investigated by Mumford *et al.* (2012).

After the values of $\tilde{b}_k(S_i)$ are estimated for all stimuli, these can be used to populate a vector $\underline{\tilde{\mathbf{b}}}_k^\top = [\tilde{b}_k(S_1), \tilde{b}_k(S_2), \ldots, \tilde{b}_k(S_{N_s})]^\top$ that describes the amplitude of the BOLD response in voxel k to all N_S stimuli used in the experiment. Similarly, after repeating this process for all voxels, we form the matrix

$$\tilde{\mathbf{B}} = [\underline{\tilde{\mathbf{b}}}_1, \underline{\tilde{\mathbf{b}}}_2, \ldots, \underline{\tilde{\mathbf{b}}}_{N_v}]. \tag{10.14}$$

The assumption that the BOLD response is proportional to aggregate neural activity means that there exists some constant λ, such that $\underline{\tilde{\mathbf{b}}}_k = \lambda \underline{\mathbf{a}}_k$ and $\tilde{\mathbf{B}} = \lambda \mathbf{A}$, where $\underline{\mathbf{a}}_k$ and $\mathbf{A}$ are the aggregate activity vector and matrix from Equations (10.12) and (10.13), respectively. Note from those equations that the voxel-based encoding model therefore predicts that

$$\underline{\tilde{\mathbf{b}}}_k = \lambda \underline{\mathbf{a}}_k = \mathbf{R}(\lambda \underline{\mathbf{w}}_k) + \lambda \underline{\boldsymbol{\epsilon}}_m \tag{10.15}$$

and

$$\tilde{\mathbf{B}} = \lambda\mathbf{A} = \mathbf{R}(\lambda\mathbf{W}) + \lambda\mathbf{E}_m. \tag{10.16}$$

Therefore, the constant λ can be absorbed into the weights and error variance. In other words, the weights and error variance include an unidentifiable constant of proportionality. This causes no problems, however, because the primary interest is not in the absolute value of the weights, but rather in their relation to each other. For example, note that if one weight in a voxel is twice as large as another weight, then this 2-to-1 ratio holds for any value of λ. As a result, without loss of generality, we can ignore λ during parameter estimation, which means that the multivariate voxel-based encoding model can be described by

$$\tilde{\mathbf{B}} = \mathbf{RW} + \mathbf{E}_m. \tag{10.17}$$

As mentioned previously, most applications either ignore channel noise or assume zero-mean, additive Gaussian noise. In either case, $\mathbf{R} = \mathrm{E}[\mathbf{R}]$, $\mathbf{E}_m$ describes the sum of channel and measurement noise, and the weight matrix $\mathbf{W}$ can be estimated from the normal equations of the multivariate version of the GLM. In most applications, the stimuli are presented far enough apart in time that it is safe to assume that the BOLD responses to separate stimuli are statistically independent. For this reason, and because it is common to assume homogeneity of variance [i.e., that each $\underline{\boldsymbol{\epsilon}}_{m,k}$ in Equation (10.13) has a multivariate normal distribution with variance–covariance matrix $\boldsymbol{\Sigma} = \sigma_m^2\mathbf{I}$], the Gauss–Markov theorem applies, and therefore the uniformly minimum variance, unbiased estimator of $\mathbf{W}$ is

$$\widehat{\mathbf{W}} = (\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\tilde{\mathbf{B}}. \tag{10.18}$$

Note that Equation (10.18) requires that $\mathbf{R}^\top\mathbf{R}$ is nonsingular. This is possible only if $N_s > N_c$, where N_s is the number of stimuli or events and N_c is the number of hypothesized channels. So the encoding model can only be tested against data in which there are more stimulus events than hypothesized channels. This makes sense, because in each voxel, there are unknown free weight parameters. To estimate these parameters uniquely, we need more data points than parameters. Each stimulus presentation produces one data point, so unique estimation of the weights requires that $N_s > N_c$. If this condition is not possible, then an alternative is to introduce extra constraints into the estimation procedure – a technique known in statistics as regularization (e.g., Bickel & Li, 2006). For example, this is the method used by Naselaris *et al.* (2009).

From a Bayesian perspective, regularization is accomplished by placing a prior on $\mathbf{W}$, so that some weight estimates are favored over others. This point is important, because regularization biases inference in favor of one $\widehat{\mathbf{W}}$ over many others that predict the same distribution of observed activity profiles $\tilde{\mathbf{B}}$. Some researchers have argued that, more than a simple technicality, this is an important theoretical decision and should be considered an important aspect of the final model (Diedrichsen, 2020; Diedrichsen & Kriegeskorte, 2017).

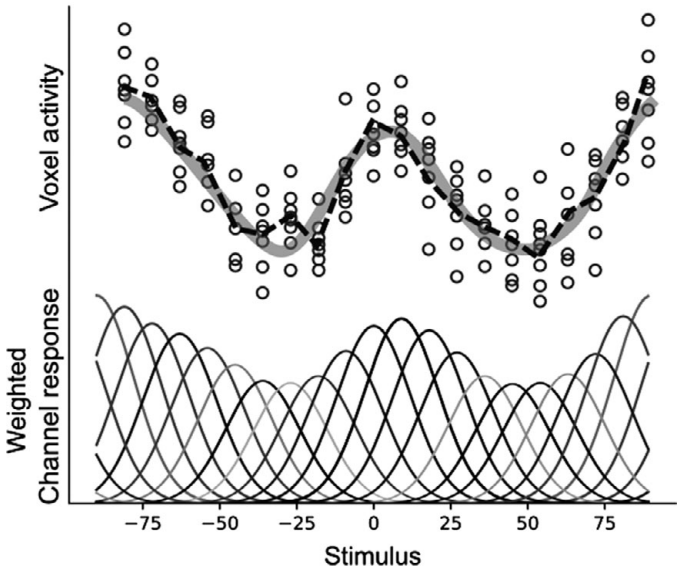


Figure 10.2 *Hypothetical data from one voxel along with theoretical predictions of the standard encoding model. Each open circle in the top half depicts a hypothetical response from this voxel on one trial of an experiment in which the stimuli vary on a single physical dimension. The scatterplot of data is called the activity profile of this voxel, and the dotted line is its mean. The channel tuning functions from the standard encoding model are shown at the bottom, each scaled by its corresponding weight parameter. The solid line in the top half is the predicted activity profile of the standard encoding model, which equals the sum of the weighted tuning functions.*

Mathematically, the combination of an encoding model for $\underline{\mathbf{r}}(S_i)$ and a linear measurement model is equivalent to regression by linear combination of basis functions (Hastie, Tibshirani, & Friedman, 2009). More specifically, the model captures the nonlinear relation between stimuli S_i and BOLD responses by using a set of nonlinear basis functions $f_c\left(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}}\right)$ to transform the stimuli, and then uses a linear model on the transformed space to predict the amplitude of the observed BOLD response $\tilde{b}_k$.

Figure 10.2 shows this more clearly with an example using the standard encoding model discussed earlier. The figure depicts hypothetical data from one voxel along with theoretical predictions of the standard encoding model that has been linked to the linear measurement model described in Equation (10.8). The hypothetical data are from an experiment in which a stimulus is presented on each trial that is a random sample from some ensemble that varies on a single physical dimension. Each open circle in the cloud of points shown in the top half of the figure depicts a hypothetical response recorded in this voxel on one trial. The value of each data point on the abscissa identifies the stimulus value on that trial. We call this scatterplot the *activity profile* of this voxel (following the nomenclature of

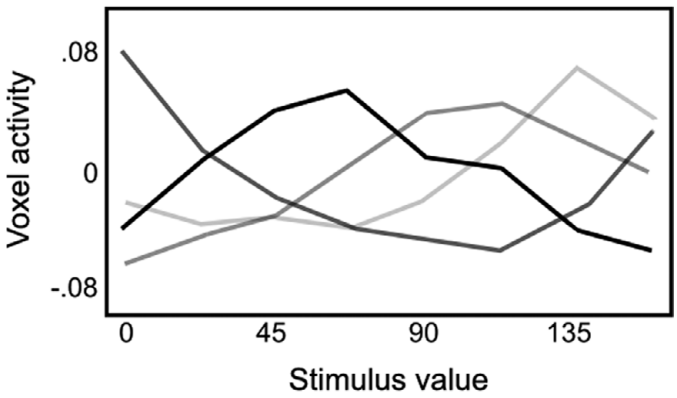


Figure 10.3 *Mean activity profiles estimated by Serences* et al. *(2009).*

Diedrichsen & Kriegeskorte, 2017), and the dotted line represents the mean of this activity profile (sometimes called the voxel tuning function). The channel tuning functions from the standard encoding model are represented at the bottom, each scaled by its corresponding weight parameter w_{ck}. So note that in this hypothetical voxel, the most under-represented channels are centered at the stimulus values -35 and $+50$. The sum of these scaled functions is represented by the solid line at the top, which accurately approximates the observed mean activity profile. In practice, the channel weights are estimated by fitting the solid-line prediction of the model to the observed data – a process known in statistics as linear regression with radial basis functions.

While more complex stimulus spaces and encoding models make the resulting model more difficult to interpret, the principle is the same: the activity profile of voxel k is modeled as a linear combination of basis functions. One issue with encoding modeling is that, in many cases, the set of basis functions will overfit the data. The reason is that the complexity and number of basis functions is selected either arbitrarily or based on theoretical considerations (e.g., the number of populations thought to underlie the voxel activity). In contrast, mean activity profiles are likely to be smooth and could probably be approximated by a small number of basis functions. For example, Figure 10.3 shows examples of mean activity profiles estimated by Serences *et al.* (2009). Note that the profiles are all unimodal and smooth, and each could probably be approximated with a single radial basis function. Although the profiles shown in this figure were averaged across many voxels, it is unlikely that much more structured variability would be found in single-voxel activity profiles, or at least not variability that can be distinguished from high levels of measurement noise common in fMRI.

Dynamic Encoding Models

All models considered so far are static, in the sense that they only predict the amplitude of the BOLD response to each stimulus. In contrast, many other

encoding models are dynamic, including, for example, dynamic causal modeling (DCM; Friston, Harrison, & Penny, 2003). These models predict changes in neural activity over time – not just because of decay in the BOLD response, but also because of dynamic changes in neural, perceptual, and cognitive processing. Dynamic encoding models require a more detailed model, not only of how neural activity changes with time, but also of how the BOLD response depends on neural activity. In particular, they require a model that predicts the entire time-course of the BOLD response, rather than just its overall amplitude.

To begin, consider the differences between static and dynamic models in their predictions about channel responses and aggregate neural activity. Many dynamic encoding models, including DCM, do not assume that aggregate neural activity is driven by a population of separate channels. Instead, in these models, aggregate neural activity is the fundamental construct. DCM compensates for this simpler account of activation within any single voxel, by using different equations to predict neural activity in different voxels – especially voxels that are in different brain regions. In contrast, voxel-based encoding models typically apply the same model (and model equations) to all voxels. The goal in this case is to identify voxels in which the observed BOLD response agrees with these predictions.

To test any encoding model against data, we first generate a predicted activity vector for each voxel in the ROI. Let the $N_{TR} \times 1$ vector $\underline{\mathbf{a}}_k^D$ denote the predicted neural activity in voxel k on every TR of the experiment. The superscript D (for dynamic) is to distinguish this vector from the static activity vector $\underline{\mathbf{a}}_k$ described in Equation (10.12). The two vectors are similar, in that they both describe aggregate activity in a voxel, but note that $\underline{\mathbf{a}}_k$ has N_S rows, whereas $\underline{\mathbf{a}}_k^D$ has N_{TR} rows. The number of TRs in an experiment cannot be less than the number of stimuli that are presented, and in most experiments N_{TR} will be much greater than N_S. Therefore, in almost all applications $\underline{\mathbf{a}}_k^D$ will have many more rows than $\underline{\mathbf{a}}_k$. Row i of $\underline{\mathbf{a}}_k$ describes the predicted aggregate activity to stimulus S_i in voxel k. In contrast, row i of $\underline{\mathbf{a}}_k^D$ describes the predicted aggregate activity in voxel k on TR i. The static vector $\underline{\mathbf{a}}_k$ includes an entry for every unique stimulus that predicts the same aggregate activity every time that stimulus is presented. The dynamic vector $\underline{\mathbf{a}}_k^D$ includes an entry that predicts the aggregate neural activity on every TR of the experiment. So if stimulus S_i is presented 10 times, then $\underline{\mathbf{a}}_k$ includes one value that predicts the same neural activity on each of these 10 presentations, whereas $\underline{\mathbf{a}}_k^D$ will predict the effects of these 10 separate presentations on every TR of the experiment.

To test a dynamic encoding model against data from multiple voxels, we first generate predicted activity vectors for each of the N_v voxels in the ROI. The next step is to form the $N_{TR} \times N_v$ activity matrix $\mathbf{A}_D$ that includes $\underline{\mathbf{a}}_{D,j}$ as column j. Note that this matrix is similar, but not identical, to the matrix $\mathbf{A}$ in Equation (10.13). They both describe aggregate activity in a set of voxels, but the columns of $\mathbf{A}_D$ are the dynamic activity vectors $\underline{\mathbf{a}}_{D,j}$, whereas the columns of $\mathbf{A}$ are the static activity vectors $\underline{\mathbf{a}}_j$.

If the model postulates an underlying population of channels that drive the aggregated neural activity, then a similar generalization is used to define the

channel responses. In particular, the model is used to form the $N_{TR} \times N_C$ channel response matrix $\mathbf{R}_D$ that contains the predicted response of channel c on every TR of the experiment in column c and the predicted response of all channels on TR i in row i. Note that the relationship between $\mathbf{R}_D$ and the static channel response matrix $\mathbf{R}$ of Equation (10.11) is similar to the relationship between $\mathbf{A}_D$ and $\mathbf{A}$. Given this dynamic channel response matrix, aggregate neural activity is predicted using a dynamic version of Equation (10.12):

$$\underline{\mathbf{a}}_k^D = \mathbf{R}_D \underline{\mathbf{w}}_k + \underline{\boldsymbol{\epsilon}}_{D,m}, \tag{10.19}$$

where the $N_{TR} \times 1$ random vector $\underline{\boldsymbol{\epsilon}}_{D,m}$ has a multivariate normal distribution with mean vector $\underline{\mathbf{0}}$ and variance–covariance matrix $\boldsymbol{\Sigma}_{D,m}$. Note that the weight vector $\underline{\mathbf{w}}_k$ is identical in the static and dynamic models. In both cases, it specifies the relative contribution of each channel to the aggregate activity. The multivoxel version of Equation (10.19) is

$$\mathbf{A}_D = \mathbf{R}_D \mathbf{W} + \mathbf{E}_{D,m}, \tag{10.20}$$

where $\mathbf{W}$ is defined exactly as in Equation (10.13).

The next problem is to model the effects of dynamic changes in aggregate neural activity on TR-by-TR changes in the BOLD response. This is a problem that has received enormous attention in the fMRI literature. Almost all current applications of fMRI assume that the transformation from neural activation to BOLD response can be modeled as a linear, time-invariant system. Although a detailed examination clearly shows that the transformation is, in fact, nonlinear (e.g., Boynton *et al.*, 1996), it also appears that the departures from linearity are not severe if the stimuli are of high contrast and brief exposure durations are avoided (Vazquez & Noll, 1998). These two conditions are commonly met in fMRI studies of high-level cognition.

Any linear, time-invariant system is completely characterized by its impulse response function, $h(t)$, which is the output of the system to an input that is an idealized impulse. More specifically, let $a(t)$ and $b(t)$ denote the (continuous-time) input and output of a dynamical system at time t, respectively. Then if the system is linear and time-invariant:

$$b(t) = a(t) * h(t) = \int_0^\infty a(\tau) h(t - \tau) d\tau, \tag{10.21}$$

for any input and for all time t (e.g., C. T. Chen, 1970).

In dynamic encoding models, the input $a(t)$ is aggregate neural activity, the output $b(t)$ is the BOLD response, and the impulse response function $h(t)$ is commonly referred to as the *hemodynamic response function* (hrf). There are a variety of different methods for selecting a functional form for the hrf (e.g., Ashby, 2019). Common choices include a gamma function or a difference of gamma functions. Some researchers have also used boxcar functions with one or more ones around the peak of the hrf and zeros elsewhere (e.g., Çukur *et al.*, 2013; Huth *et al.*, 2012; Nishimoto *et al.*, 2011). In all cases, however, parameters are chosen so that the resulting hrf peaks at around 6 s and then decays back to 0 after 30 s or so.

Dynamic encoding models make dynamic predictions about how neural activation $a(t)$ changes moment-by-moment. Therefore, in such models, Equation (10.21) is used to convert model predictions to values of the observed dependent variable – that is, to values of the BOLD response $b(t)$.

Equation (10.21) assumes that the BOLD response is measured in continuous time. In practice, however, the BOLD response is measured only at discrete time points separated by a fixed amount of time equal to the TR. So rather than a continuous-time integral, the Equation (10.21) convolution is done in discrete time. This can be accomplished using simple matrix multiplication by loading values of the hrf into the appropriate Toeplitz matrix.[1]

The Toeplitz matrix, which has order $N_{TR} \times N_{TR}$, includes a time-lagged discrete representation of the hrf in each column. To build the matrix, we begin by discretizing the hrf in a way that is similar to how we discretized the neural predictions of the model. The only difference is that any reasonable model of the hrf will include nonzero values only for 30 s or so, whereas the functional run is likely to last 5 min or longer. Suppose we assume that the BOLD response to an impulse of neural activation persists for at most N_h TRs (since the hrf is an impulse response function). Then our discretized version of the hrf will be a vector $\underline{\mathbf{h}}^\top = \left[h_1, h_2, \ldots, h_{N_h}\right]^\top$, where $h_i = h(t = i \times TR)$. Next, we use $\underline{\mathbf{h}}$ to build the appropriate Toeplitz matrix:

$$\mathbf{H} = \begin{bmatrix} h_1 & 0 & 0 & \ldots & 0 \\ h_2 & h_1 & 0 & \ldots & 0 \\ h_3 & h_2 & h_1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ h_{N_h} & h_{N_h-1} & h_{N_h-2} & \ldots & 0 \\ 0 & h_{N_h} & h_{N_h-1} & \ldots & 0 \\ 0 & 0 & h_{N_h} & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & h_{N_h} \end{bmatrix}. \tag{10.22}$$

Given this matrix, the discrete-time version of the Equation (10.21) integral reduces to the simple matrix multiplication

$$\underline{\mathbf{b}} = a\,(t) * h\,(t) = \mathbf{H}\underline{\mathbf{a}}_D. \tag{10.23}$$

Therefore, note that the dynamic encoding model predicts that the observed BOLD response in voxel k on each TR equals $\underline{\mathbf{b}}_k = \mathbf{H}\underline{\mathbf{a}}_{D,k}$.

The dynamic version of the voxel-based encoding model, which assumes that aggregate activity is driven by a population of channels, is generalized from Equation (10.20) by noting that the predicted aggregate activity matrix $\mathbf{A}_D = \mathbf{R}_D\mathbf{W}$ and therefore the predicted $N_{TR} \times N_v$ BOLD response matrix $\mathbf{B} = \mathbf{H}\mathbf{A}_D$. Combining these produces the multivoxel, dynamic voxel-based encoding model

1 A Toeplitz matrix is any matrix in which all descending diagonals are filled with the same value.

$$\mathbf{B} = \mathbf{H}\mathbf{R}_D\mathbf{W} + \mathbf{E}_D, \tag{10.24}$$

where $\mathbf{E}_D$ is now a combination of noise at the level of neural channels, voxel activities, and BOLD responses.

The traditional GLM analysis of fMRI data, which is typically implemented in the popular fMRI data analysis software packages SPM and FSL, can be considered a special case of Equation (10.24) (van Gerven, 2017), in which different channels respond to different stimulus events (e.g., each different type of stimulus, the participant's response, feedback, etc.), and each channel response is a boxcar function of zeros and ones, representing the absence and presence, respectively, of that event on each TR. Therefore, the true contribution of encoding models is not in the linearized measurement model, which was already available in the standard GLM approach, but rather in the much more detailed models of the possible computations performed by each channel.

The models we have considered so far all assume that the transformation from neural activity to BOLD response can be modeled as a linear, time-invariant system. More detailed models attempt to account for nonlinearities in the BOLD response. The most popular is the balloon model (Buxton, Wong, & Frank, 1998), which models key biomechanical properties of the brain's vasculature. The balloon model accounts for the conflicting effects of dynamic changes in both blood oxygenation and blood volume and assumes that the blood flow out of the system depends on a balloon-like pressure within the vasculature. For example, when the blood flow is high, the walls of the blood vessels are under greater tension, and as a result they push the blood out with greater force, which reduces the rate at which oxygen is extracted from the hemoglobin. DCM, as implemented in the fMRI software package SPM (i.e., DCM10/SPM8), converts predicted neural activations to BOLD responses via a generalization of the balloon model. In contrast, most encoding models settle for a linear systems approach, and therefore instead convert predicted neural activations to BOLD responses via the convolution integral of Equation (10.21).

10.2.3 Population Receptive Fields

The population receptive field (pRF) of a voxel is a description of the region of the visual field where stimulus presentations produce an fMRI response (Dumoulin & Wandell, 2008; Wandell & Winawer, 2015). For example, panel (a) in the right column of Figure 10.4 shows the pRF of the traditional approach, which assumes that the pRFs of all individual neurons in a voxel can be described by a single population-level pRF.

In its traditional implementation, the presented stimulus is represented by an indicator function $s(x, y) = \{0, 1\}$, where the values 0 and 1 denote the absence and presence, respectively, of any part of a stimulus at spatial coordinates (x, y). The pRF is modeled by a two-dimensional isotropic Gaussian in the same space:

$$g(x, y) = \exp\left[\frac{(x - x_0)^2 + (y - y_0)^2}{2\varsigma^2}\right], \tag{10.25}$$

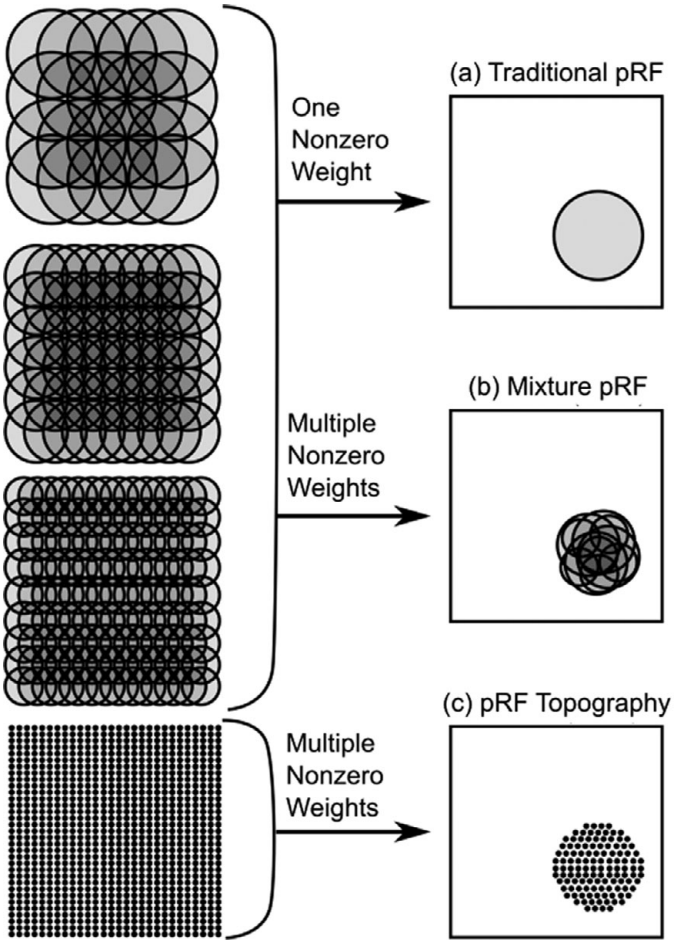


Figure 10.4 *The population receptive field (pRF) method can be seen as an application of encoding modeling. (a) In the traditional approach, the pRFs of all individual neurons in a voxel can be described by a single population-level pRF (Dumoulin & Wandell, 2008). (b) Mixture pRFs assume the voxel includes channels with different receptive fields (Sprague & Serences, 2013). (c) pRF topography assumes that the receptive field of each channel in the voxel is a Kronecker delta function (Lee* et al.*, 2013).*

where (x_0, y_0) is the center (i.e., mean) and ς the spread (i.e., standard deviation) of the receptive field. The predicted response of a voxel in which the pRFs of all neurons can be described by this single population-level pRF is computed by location-by-location multiplication of the stimulus value and the voxel pRF and then summing all these responses:

$$r(s_i) = \int_{x_L}^{x_U} \int_{y_L}^{y_U} s_i(x,y)g(x,y)dxdy, \tag{10.26}$$

where x_L, x_U, y_L, and y_U represent the lower (L) and upper (U) boundaries of the visual field along the x and y coordinates. As in most applications, the model

implicitly assumes that $r(s_i)$ includes additive Gaussian neural noise. The voxel activity is assumed to be a scaled version of the population response

$$a_k(s_i) = w\, r(s_i), \tag{10.27}$$

and the BOLD response is modeled as described in the previous section. Estimating the pRF of a voxel is done by finding the values of the parameters x_0, y_0, and ς that allow the model to provide the best possible fit to the observed BOLD response.

The pRF technique is usually considered an alternative to the linearized encoding modeling that is the focus of this chapter, but it can also be seen as a special case of the general encoding model framework. As shown in Figure 10.4a, the problem of estimating a pRF can be recast as an encoding modeling problem. First, one creates an encoding model with a large number of channels, each having a receptive field with a slightly different position and size, as illustrated in the left column of Figure 10.4. Second, to mimic the traditional pRF approach, one constrains all channel weights to be zero except for one, in order to accommodate the assumption that the pRFs of all neurons in the voxel can be modeled by one population-level pRF, and therefore that the data from each voxel can be modeled by a single channel. The traditional pRF approach is therefore equivalent to assuming a large number of channels that densely cover the space of possible size and location parameters, and then finding the single nonzero weight that provides the best fit to the data. The single channel with a nonzero weight has the position and size of the traditional pRF.

Of course, a more traditional encoding model that includes many channels also could be used to describe the pRF (see Figure 10.4b). This model would include nonzero weights for multiple channels, with each channel characterized by a receptive field of different position and size. Sprague and Serences (2013) used such a mixture model to study the effects of spatial attention on neural representations in visual cortex. After the model is fitted to data, the pRF is equivalent to the predicted mean activity profile (the solid line curve in the top half of Figure 10.2). The resulting pRF is likely to be similar to the one obtained by assuming a single channel, but this encoding modeling approach has the advantage of more transparently reflecting the empirical observation that the voxel pRF is a mixture of multiple neural receptive fields of smaller size (Dumoulin & Wandell, 2008).

Other advantages of describing pRFs as applications of encoding modeling are that it encompasses other techniques proposed to obtain pRFs, it facilitates the understanding of how different techniques relate to one another, and suggests new techniques that could be useful in research. Because the linearized encoding model can be understood as linear regression with basis functions, alternative pRF models are easily obtained simply by changing the basis functions or the constraints used to estimate weights. For example, Lee *et al.* (2013) proposed an alternative method for estimating pRFs, illustrated in Figure 10.4c, which uses Kronecker

delta functions (i.e., impulses) as the basis set. In this approach, the pattern of estimated weights directly models the pRF topography.

Insights obtained from the pRF approach could also benefit encoding modeling more generally. In particular, pRFs are defined in the stimulus space and their parameters have interpretable units, which allows researchers to make meaningful comparisons across participants, conditions, and measurement instruments (Wandell & Winawer, 2015). As discussed in the next section, the parameters of a fitted encoding model can be difficult to interpret. The pRF approach, however, allows researchers instead to focus on characterizing, for each voxel, the mean activity profile predicted by a fitted encoding model (solid curve in the top half of Figure 10.2). Most commonly this means estimating the mode and spread of the mean activity profile, but other features of the function (support, derivatives, etc.) may also be informative. Unlike the traditional pRF approach, an encoding modeling approach could describe selectivity along any stimulus dimension, not only spatial sensitivity within visual field space.

10.2.4 Feature Spaces and Model Interpretation

The development so far is quite general, in the sense that it encompasses voxel-based encoding models, the standard GLM approach to constructing a statistical parametric map, population receptive fields, and model-based fMRI (developed in more detail below). What is common to these different approaches is the use of a linear measurement model with Gaussian noise (i.e., the GLM). They differ mainly in how they define a channel and a channel response [i.e., $r_c(S_i)$]. The space of channel responses is sometimes called feature space (e.g., Diedrichsen, 2020), and the power and flexibility of the encoding modeling approach lies in the possibility of choosing among many different feature spaces.

For example, Naselaris *et al.* (2009) constructed one voxel-based encoding model in which each channel was a Gabor wavelet and another in which each channel responded to a different semantic category of objects – for example, birds, fish, or vehicles. Whereas the Gabor wavelet encoding model gave good accounts of BOLD responses in low-level visual cortical areas, the semantic encoding model gave good accounts in high-level association areas. So an encoding model approach can be used to identify brain regions that are sensitive to whatever features are hypothesized to drive the channel responses. A model based on features that do not match any set of channels in the brain should provide a poor account of BOLD responses in all ROIs.

The Gabor wavelet model was motivated by a long line of vision research on the sensitivity of V1 neurons to spatial frequency and orientation. In the case of high-level visual areas, however, the decision about how to define the channels is often more arbitrary. For example, in the case of the semantic-encoding model, the decision was made to include channels that respond to the presence of certain categories of natural objects (e.g., birds and fish), but not others, and the object classes that were chosen had to be hand coded in every image by human observers

(e.g., does this image contain a bird?). More recently, there have been a number of attempts to identify features, and therefore to define the underlying channels, by using artificial neural networks (e.g., Eickenberg *et al.*, 2017; Güçlü & van Gerven, 2015). The general approach is to construct a multilayer neural network – commonly a deep convolutional neural network – and then train it to classify a database of natural images. After training, the output of each layer is interpreted as a different possible set of channel responses, and these are compared to the BOLD responses from different ROIs within the visual system.

For example, Güçlü and van Gerven (2015) trained a deep neural network that included five convolutional and three fully connected layers to classify images into 1 of 1,000 different object categories. The network was trained on a database of around 1.2 million natural images using a supervised learning algorithm. After training was complete, each of the eight layers of the network were used to define eight different possible sets of channels, and therefore eight different encoding models. Each of these eight models was then tested against the fMRI data reported by Naselaris *et al.* (2009) by using an output model similar to Equation (10.12). Overall, the models gave good accounts of visual responses across the entire ventral stream. Furthermore, the BOLD responses in early visual areas were best accounted for by early network layers, whereas in higher-level (i.e., downstream) visual areas, the BOLD responses were best fit by higher-level network layers.

The neural network used in this application included some features that were inspired by neural processing in the human brain (e.g., convolutional layers). But the model has much closer ties to the machine-learning literature than to neuroscience. Essentially, it can be viewed as an attempt to build an optimal model of object classification. The fact that it gives a good account of BOLD responses in visual cortex as humans view images of natural scenes suggests that the human visual system may have evolved to optimize object classification.

In sum, the feature space can be the response of filters to images, the responses of units in a deep neural network, variables in an abstract cognitive model, labels applied by researchers to their stimuli, etc. This flexibility allows researchers to propose multiple competing feature spaces to explain neural activity in a particular brain region, and use model selection techniques (Zucchini, 2000) to choose one that describes the data best without overfitting. Ideally, the set of competing models would include only feature spaces that are theoretically relevant, preferably supported by evidence from past research.

Unfortunately, the flexibility and power of encoding models also leads to a number of issues of model interpretation. The first problem is that sometimes it is unclear whether the feature space is a representation of the stimuli S_i or of the neural channel responses r_c. Many encoding models provide a separate notation for stimuli and channel responses, together with equations indicating how to compute channel responses given the presentation of a stimulus. On the other hand, some applications have used a set of hand-coded stimulus labels as the feature set (e.g., Çukur *et al.*, 2013; Huth *et al.*, 2012), with binary indicator

variables used to represent such labels. In this case, it is unclear whether those variables are assumed to represent the presence of a stimulus or the response of a channel that is dedicated to the detection of that stimulus. If one assumes that the feature space is a representation of the stimuli, then the linear measurement model assumes a linear mapping, not only from channels to measurements, but also from stimuli to neural responses. Both would be described by the estimated parameter matrix $\widehat{\mathbf{W}}$ [i.e., from Equation (10.18)]. On the other hand, if one assumes that the labels are a representation of channel responses (i.e., populations of neurons that are active when the stimulus feature is presented), then there is an unknown transformation between S_i and r_c, which is likely nonlinear and is not explicitly modeled. The way in which most researchers discuss their results suggests that the latter interpretation is most common. For example, when Naselaris *et al.* (2009) compared the Gabor wavelet model against the semantic model that was constructed by hand coding labels in each image, they implicitly assumed that both models were identical except for the type of features to which the underlying channels were tuned. What this type of comparison does not take into account is the quality of the encoding models themselves. For example, only the Gabor wavelet model provides an explicit mechanistic description of how each channel responds to any possible stimulus.

The second issue has to do with the interpretation of the weight matrix $\widehat{\mathbf{W}}$. It is tempting to interpret estimated weights as providing information about the relative importance of different channels in the activity of a given voxel. This was the interpretation we assigned each weight when building the model [e.g., see Equation (10.8)]. However, those were forward inferences, whereas interpreting entries in $\widehat{\mathbf{W}}$ after model fitting is a backward inference. And in the case of encoding models at least, backward inferences are tricky. There are multiple reasons why the entries in $\widehat{\mathbf{W}}$ might not provide the expected weight information (Kriegeskorte & Douglas, 2019). For example, in most cases, channels are not chosen to provide responses that are independent of each other, so multicollinearity among the channel responses may occur. Under these circumstances, weights are difficult to interpret because they do not reflect the effect of each channel independently from all others. In addition, some models are over-parameterized, in the sense that many different weight matrices describe the data equally well [i.e., so $\mathbf{R}^\top\mathbf{R}$ in Equation (10.18) is singular]. In practice, such identifiability problems are solved using regularization, but this reflects the choice of a particular prior over weights (Diedrichsen & Kriegeskorte, 2017). A channel with a large weight under one prior could have no weight under a different prior, so interpretation of weights should take into account what assumptions about the measurement model are implemented by the chosen prior.

A third, related issue has to do with interpreting the success of an encoding model to describe data from a given voxel as evidence that the feature space of the model is represented in the voxel. This is called the feature fallacy error because, for any given feature space used to describe voxel activities, there are an infinite number of other feature spaces that will make the exact same predictions,

given that the matrix of weights $\widehat{\mathbf{W}}$ is modified accordingly (e.g., by choice of an appropriate prior; Diedrichsen, 2020; Diedrichsen & Kriegeskorte, 2017).

Gardner and Liu (2019) recently showed why this is the case for the standard linearized encoding model described by Equation (10.13). For example, consider a model, call it Model 1, in which the predicted activity matrix $\mathbf{A}$ equals

$$\mathbf{A} = \mathbf{R}_1\mathbf{W}_1, \tag{10.28}$$

where $\mathbf{R}_1$ is the expected value of the channel response matrix. Now consider a second model, Model 2, that postulates a different set of expected channel responses $\mathbf{R}_2$ that are linearly related to the Model 1 responses via

$$\mathbf{R}_2 = \mathbf{R}_1\mathbf{P}, \tag{10.29}$$

where $\mathbf{P}$ is some $N_c \times N_c$ nonsingular matrix. Therefore, note that the predicted aggregated activity matrix for Model 2 is

$$\mathbf{A}_2 = \mathbf{R}_2\mathbf{W}_2 = \mathbf{R}_1\mathbf{P}\mathbf{W}_2. \tag{10.30}$$

Now if $\mathbf{W}_2 = \mathbf{P}^{-1}\mathbf{W}_1$, it follows that

$$\mathbf{A}_2 = \mathbf{R}_1\mathbf{P}\mathbf{P}^{-1}\mathbf{W}_1 = \mathbf{R}_1\mathbf{W}_1 = \mathbf{A}_1, \tag{10.31}$$

and therefore, both models predict exactly the same aggregated activity matrix, even though they postulate different channel responses and different weights. Diedrichsen and Kriegeskorte (2017) argued that similar model identifiability problems arise even when weights are estimated using regularization rather than by solving the normal equations [as in Equation (10.18)].

The identifiability and model mimicry problems that are endemic to encoding models are likely not restricted to models that span the exact same linear subspace. This becomes clear if we refer back to the Figure 10.2 example, which we used to illustrate that encoding models are a form of linear regression with radial basis functions. The radial basis functions illustrated in the bottom part of Figure 10.2 are not the only ones that could provide a good fit to the activity profile shown in the top part of the figure. Given enough channels, a model in which the basis functions are polynomials, splines, or even simple step functions could provide an arbitrarily good fit (see Hastie, Tibshirani, & Friedman, 2009).

What all this means is that one must be extremely careful when interpreting the success of an encoding model in terms of its basis functions or features. Sometimes a particular set of features is theoretically important, neurobiologically motivated, or simply easier to interpret. All of these are good reasons to prefer one basis set over others. At the same time, however, it is essential to acknowledge that the fit and predictive performance of a model do not guarantee, by themselves, that an ROI encodes stimuli using that specific basis set.

10.3 Model Inversion

Although encoding models provide the best opportunity to make causal inferences from fMRI data (Weichwald *et al.*, 2015), decoding methods offer their own distinct advantages (e.g., Naselaris *et al.*, 2011). One is that they allow decoding accuracy to be compared directly to human behavioral performance in each ROI. For example, D. B. Walther *et al.* (2009) compared the confusions that human observers made when categorizing natural scenes with the confusions made by an MVPA classifier in a variety of different visual ROIs. Although the human observers made fewer errors, the pattern of confusions made by the MVPA classifier in the parahippocampal place area was similar to the pattern of confusions made by the humans, whereas the pattern of confusions made by the classifier in V1 was not correlated (at least, not significantly) with the pattern made by the humans. Thus, this result supports a model in which the parahippocampal place area plays a key role in scene classification behavior.

Carlson and colleagues extended this approach by assuming that the observer's response time on each trial is related to the distance of the activity pattern to the best-fitting linear bound of an MVPA classifier (Carlson *et al.*, 2013; Grootswagers, Cichy, & Carlson, 2018; Ritchie & Carlson, 2016; Ritchie, Tovar, & Carlson, 2015). The assumption that response time is inversely related to the distance between the percept and a decision bound has a long history in mathematical psychology (e.g., Ashby & Maddox, 1994; Murdock, 1985; Chapter 6 of this volume). Thus, if a particular brain region stores information that is extracted for behavioral performance, then it is likely that distances-to-bound obtained from a classifier trained on data from that region will correlate with response times and similar behavioral measures. Using this approach, the Carlson group has shown that brain regions that provide information that is read out for behavior are only a subset of the brain regions that contain decodable information.

Decoding methods are also popular because they provide the basis of the popular claims that fMRI can be used for mind reading (Haynes & Rees, 2006). In these applications, the BOLD responses are decoded to predict the stimulus event that occurred. Many exciting possibilities have been proposed – from communicating with patients who were diagnosed to be in a vegetative state, to lie detection, to enabling people to control external devices via thought (DeCharms, 2008).

Researchers who develop and test encoding models can exploit many of the advantages of decoding approaches via model inversion, which is the process of constructing a decoding scheme by inverting an encoding model. Perhaps the most immediate benefit of this process is that it allows unique tests of the encoding model that would otherwise be impossible. For example, a valid encoding model that accurately predicts how the BOLD response differs when different stimuli are presented should also be able to predict which stimulus was presented simply by examining the BOLD response on each trial. In mathematical psychology, the validity of a model is typically assessed by examining its ability to predict what response was made (and perhaps also the response time), given knowledge of the

stimulus. An inverted encoding model allows tests in the opposite direction – that is, it allows a test of the model's ability to predict what stimulus was presented, given knowledge of the response.

An encoding model predicts the aggregate activity in a voxel given knowledge of the stimulus [e.g., see Equation (10.8)]. More specifically, a complete encoding model should predict the probability density function of aggregate activity in voxel k on trials when stimulus S_i is presented – that is, $P(a_k|S_i)$. In this approach, Bayes' rule is used to invert the model:

$$P(S_i|a_k) \propto P(a_k|S_i)P(S_i), \tag{10.32}$$

where $P(S_i)$ is the prior probability that stimulus S_i is presented. When stimuli are modeled in a physical stimulus space, such as the pixel space used to construct each stimulus, model inversion allows for full reconstruction of the presented stimulus. Of course, decoding is possible without the use of an explicit encoding model, as in MVPA, by training machine-learning algorithms to extract information about stimuli from activity patterns (see Pereira, Mitchell, & Botvinick, 2009).

The Equation (10.32) decoding scheme operates directly on the model's predicted aggregate activities. As we saw earlier, however, many models predict that aggregate activity is determined by the responses of a population of underlying channels [e.g., as in Equation (10.8)]. These hypothesized channels have important consequences for model inversion. In particular, in addition to using the observed BOLD response to make inferences about what stimulus was presented (i.e., stimulus decoding), model inversion often makes it possible to use the observed BOLD response to make inferences about the channel responses, which typically are unobservable. Estimating the channel responses from a decoding scheme is a form of *population response reconstruction*.

Of course, if the channel responses are observable, then they could also be used for stimulus decoding. In other words, one could predict the presented stimulus either from the aggregated activity (i.e., the BOLD response) or from the channel responses. It is very important, however, to keep the distinction between these two forms of stimulus decoding in mind when interpreting the results of encoding and decoding studies. For example, the act of perception is a form of stimulus decoding because the brain must use neural activity to make inferences about the presented stimulus. But this decoding process must use channel responses. In fMRI experiments, the aggregate activity is the total neural activity in tens of thousands of neurons located in an arbitrarily defined cube of the brain. The neurons in this cube likely project to a variety of different targets, and therefore the downstream neurons are driven by the channels, not by the aggregated activity. Conversely, note that the fMRI experimenter has indirect access to the aggregated activity (i.e., via the BOLD response), but typically has no access to the responses of individual channels. Therefore, whereas the brain can only decode the stimulus from the channel responses, the experimenter can only decode the stimulus from the aggregated activity. Despite this important difference, it is common to find conflation of $\mathbf{r}$ (the channel response to stimulus S_i) and $\mathbf{a}_i$ (the aggregate activity

in response to stimulus S_i; e.g., Bobadilla-Suarez *et al.*, 2020; Diedrichsen & Kriegeskorte, 2017), which may lead to incorrect theoretical conclusions.

During model inversion, researchers usually distinguish between training and testing data. The standard approach is to first use a set of training data from some ROI to fit the encoding model (i.e., estimate all free parameters). Next, the encoding model is inverted to create a decoding scheme. Finally, the decoding method is tested against new validation data from the same ROI.

To begin, let $\tilde{\mathbf{B}}_{\text{train}}$ and $\tilde{\mathbf{B}}_{\text{test}}$ denote the data matrices collected in the ROI during training and testing, respectively. Both matrices have order $N_s \times N_v$ and, as described by Equation (10.14), they contain the amplitude of the BOLD response to all N_S stimuli in all N_v voxels. Row i summarizes the BOLD response to stimulus S_i in every voxel, and column k summarizes the response in voxel k to every stimulus. Now consider encoding models in which aggregate activity is assumed to depend on responses from an underlying population of channels. In these models, the channel-response matrix $\mathbf{R}$ depends on exactly which stimuli are presented and on their order of presentation. The training and testing data might come from trials that present the same stimuli, but even in this case the order of stimulus presentation will typically differ. Therefore the channel-response matrices for training and testing will differ. Denote these two matrices by $\mathbf{R}_{\text{train}}$ and $\mathbf{R}_{\text{test}}$, respectively. Although encoding models assume the expected values of these two matrices will differ, they assume that the matrix of channel weights $\mathbf{W}$ will be the same during training and testing. This is because $\mathbf{W}$ depends on the relative frequencies of the different channels in the voxels within the search set, but not on the stimuli that are presented [i.e., see Equation (10.13)].

10.3.1 Population Response Reconstruction

Given that the population responses of the hypothesized channels are not directly observable with fMRI, an interesting application of model inversion is to estimate these responses (Brouwer & Heeger, 2009). In fact, this one application is what researchers in the literature usually refer to as "inverted encoding modeling" or IEM (e.g., Gardner & Liu, 2019; Liu, Cable, & Gardner, 2018; Sprague, Boynton, & Serences, 2019; Sprague *et al.*, 2018).

According to the multivariate encoding model described in Equation (10.17), the predicted (i.e., mean) BOLD amplitude during training is $\widehat{\tilde{\mathbf{B}}}_{\text{train}} = \text{E}[\mathbf{R}_{\text{train}}]\widehat{\mathbf{W}}$. Note that $\widehat{\tilde{\mathbf{B}}}_{\text{train}}$ and $\tilde{\mathbf{B}}_{\text{train}}$ are different. The former is the predicted BOLD response according to the model, whereas the latter is the observed BOLD response. Now to fit the encoding model to the training data, we first compute $\text{E}[\mathbf{R}_{\text{train}}]$ from the model, and then use $\tilde{\mathbf{B}}_{\text{train}}$ to compute $\widehat{\mathbf{W}}$ [from Equation (10.18)]. Our goal now is to use the $\widehat{\mathbf{W}}$ matrix we estimated from the training data and the observed voxel activities during testing (i.e., $\tilde{\mathbf{B}}_{\text{test}}$) to estimate the matrix of expected population responses $\text{E}[\mathbf{R}_{\text{test}}]$, which we abbreviate as $\hat{\mathbf{R}}_{\text{test}}$. If we know these channel responses then we can infer which stimulus was presented simply by

comparing the estimated channel responses (i.e., the rows of $\widehat{\mathbf{R}}_{\text{test}}$) to each row of the original expected channel-response matrix E[$\mathbf{R}_{\text{train}}$] [see Equation (10.14)] – assuming that the stimuli presented during testing were all presented one or more times during training.

At testing, the encoding model predicts that the BOLD responses should be

$$\widehat{\tilde{\mathbf{B}}}_{\text{test}} = \widehat{\mathbf{R}}_{\text{test}}\widehat{\mathbf{W}}. \tag{10.33}$$

Our goal is to solve for $\widehat{\mathbf{R}}_{\text{test}}$. Unfortunately, however, since at this stage of the analysis $\widehat{\mathbf{R}}_{\text{test}}$ is unknown, so is $\widehat{\tilde{\mathbf{B}}}_{\text{test}}$. If we did know $\widehat{\tilde{\mathbf{B}}}_{\text{test}}$, then we could just solve for $\widehat{\mathbf{R}}_{\text{test}}$. Ester, Sprague, and Serences (2015) proposed estimating $\widehat{\tilde{\mathbf{B}}}_{\text{test}}$ with the observed data $\tilde{\mathbf{B}}_{\text{test}}$, and then solving the resulting equation for $\widehat{\mathbf{R}}_{\text{test}}$. This process produces the following estimator:[2]

$$\widehat{\mathbf{R}}_{\text{test}} = \widehat{\tilde{\mathbf{B}}}_{\text{test}}\widehat{\mathbf{W}}^{\top}\left(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top}\right)^{-1}. \tag{10.37}$$

Note that $\widehat{\mathbf{W}}$ has order $N_c \times N_v$, so $\left(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top}\right)^{-1}$ exists only if $N_v \geq N_c$ – that is, only if there are at least as many voxels in the ROI or searchlight as there are channels. Adding more voxels to the ROI adds more data (i.e., each new voxel adds a column to $\mathbf{B}$), but the size of the search volume does not affect the size of $\mathbf{R}$ (since $\mathbf{R}$ has order $N_s \times N_c$). So the more voxels there are in the search volume, the more data we have to estimate the rows of E[$\mathbf{R}_{test}$].

As an example of how Equation (10.37) is applied, Ester, Sprague, and Serences (2015) used this approach to study visual representations during the delay period of a working-memory task in which subjects had to remember the orientation of a briefly presented Gabor pattern. The encoding model assumed nine different orientation channels. They used a leave-one-run-out cross-validation procedure (e.g., see Ashby, 2019) in which they fit the encoding model to the data from all but one functional run by estimating the weight matrix $\mathbf{W}$ from these data using Equation (10.18). Next, they used the data from the single withheld functional run to invert the encoding model – that is, to estimate E[$\mathbf{R}_{test}$] from Equation (10.37), which provided an estimate of the channel responses during the delay period of each trial of the withheld run. In brain regions that maintain a visual representation of the stimulus during the delay period, the estimated channel responses should peak at the to-be-remembered orientation, whereas in any other region, the channel

2 If we estimate the predicted matrix $\widehat{\tilde{\mathbf{B}}}_{\text{test}}$ with the observed data matrix $\tilde{\mathbf{B}}_{\text{test}}$, then Equation (10.33) becomes

$$\tilde{\mathbf{B}}_{\text{test}} = \widehat{\mathbf{R}}_{\text{test}}\widehat{\mathbf{W}}. \tag{10.34}$$

Multiplying both sides by $\widehat{\mathbf{W}}^{\top}(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top})^{-1}$ produces

$$\tilde{\mathbf{B}}_{\text{test}}[\widehat{\mathbf{W}}^{\top}(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top})^{-1}] = \widehat{\mathbf{R}}_{\text{test}}\widehat{\mathbf{W}}[\widehat{\mathbf{W}}^{\top}(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top})^{-1}], \tag{10.35}$$

which implies

$$\widehat{\mathbf{R}}_{\text{test}}(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top})(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top})^{-1} = \tilde{\mathbf{B}}_{\text{test}}\widehat{\mathbf{W}}^{\top}(\widehat{\mathbf{W}}\widehat{\mathbf{W}}^{\top})^{-1}, \tag{10.36}$$

from which Equation (10.37) easily follows.

responses should all be roughly the same. Using this approach, Ester, Sprague, and Serences (2015) were able to identify a broad network of frontal, parietal, and occipital regions that maintained a high-fidelity visual representation during the delay period.

This method has also been used to study how psychological factors such as attention (Garcia, Srinivasan, & Serences, 2013; Sprague & Serences, 2013), working memory (Ester *et al.*, 2013), or learning (Byers & Serences, 2014; Ester, Sprague, & Serences, 2020) influence population responses. In these studies, $\widehat{\mathbf{W}}$ is estimated from training data, and then separate population responses are estimated from data collected in two or more test conditions [using Equation (10.37)], each run under different levels of the psychological factor (e.g., with and without attention). Finally, these separate estimates are all compared.

Recall that row i of $\mathbf{R}$ lists the response of each channel in the population to presentation of stimulus S_i. If the tuning functions all have the same shape during both training and testing (i.e., the model is homogeneous), then each row of $\mathbf{R}$ should peak at the channel most sensitive to S_i and then decay as predicted by the channel tuning function $f_c(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}})$ [i.e., see Equation (10.1)]. To estimate this function, it is common to shift the rows in $\widehat{\mathbf{R}}_{\text{test}}$ so that the peak of the response is in the same place across all channels (this is usually facilitated by the use of circular dimensions, such as orientation or color), followed by averaging of responses across rows. However, this method will fail if tuning is not homogeneous, which could happen, for instance, if the test condition influences some channels more than others (Hays & Soto, 2020).

There has been much recent controversy regarding the correct interpretation of population responses that are estimated by inverting an encoding model (e.g., Gardner & Liu, 2019; Liu, Cable, & Gardner, 2018; Sprague *et al.*, 2018). What does it mean to find, for example, that attention narrows the estimated population responses, or that it increases their amplitude? When the standard encoding model is assumed, a change in the channel tuning function $f_c(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}})$ produces a corresponding change in the population responses. However, the converse is not necessarily true: if Equation (10.37) is used to estimate the population responses, then a change in those estimates across conditions does not imply a corresponding change in the channel tuning functions.

For example, Liu, Cable, and Gardner (2018) reported evidence that the Equation (10.37) estimates of the population responses can be biased by noise. They ran an experiment in which gratings were presented at one of two different contrasts. Single-unit electrophysiology shows that orientation tuning is contrast invariant, so the width of the orientation channels should be the same for the two contrasts. Therefore, the population responses estimated via Equation (10.37) should be contrast invariant. In violation of this prediction, Liu, Cable, and Gardner (2018) found that the estimated population response widths were greater for the low-contrast gratings and they reported results of simulations supporting the hypothesis that this apparent bias was the result of decrements in signal-to-noise ratio that occur when contrast is reduced.

Sprague *et al.* (2018) defended the inverted encoding model approach of Equation (10.37) by correctly pointing out that its goal is to make inferences about population responses $\mathbf{r}$, not about individual tuning functions $f_c(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}})$. However, there seems to be a lack of clarity regarding the correct interpretation of an estimated population response. In terms of brain processing, channel responses are important because they are the input for downstream neurons that are part of the decoding network that makes perception possible (and more generally, any behavior). Any narrowing of the tuning function that might be caused, for example, by attention, therefore provides more precise downstream information for decoding. For this reason, Liu, Cable, and Gardner (2018) are also correct when they point out that the information available for stimulus decoding is better characterized by the posterior distribution over stimuli $P(S_i|a_k)$ [i.e., see Equation (10.32)] than by any reference to population responses (Van Bergen *et al.*, 2015).

A focus on $P(S_i|a_k)$ would also avoid a common issue in the literature, which is that many researchers interpret estimated population responses by reference and comparison to tuning functions from single-cell recordings, rather than by focusing on what population responses would mean for downstream processing. This is likely the result of how foreign the concept of a population response is to an experimental neuroscientist. Electrophysiologists rarely measure the response of multiple neurons or populations to a single stimulus. Instead, they typically measure the response of a single neuron or small number of neurons to many stimuli. For this reason, when Sprague *et al.* (2018) discuss population responses, a casual reader could misinterpret their use of "population-level channel response functions" as something like the channel tuning functions $f_c(S_i, \underline{\boldsymbol{\theta}}_c, \underline{\mathbf{x}})$, rather than their intended meaning as a pattern of distributed activity across channels (i.e., $\mathbf{r}$).

On the other hand, a focus on $P(S_i|a_k)$ does not solve all the issues with model inversion highlighted by the Liu, Cable, and Gardner (2018) results. In particular, inversion of an encoding model that does not capture some of the data-generating mechanisms will often lead to the wrong conclusions. In the Liu, Cable, and Gardner (2018) study, the mechanism left out of the model was the influence of contrast on signal-to-noise ratio. Unfortunately, whether one inverts the model to obtain estimates of $\mathbf{r}$ or $P(S_i|a_k)$, such estimates will be biased when the encoding model is grossly incorrect.

Although Equation (10.37) provides biased estimates of channel tuning functions, it nevertheless is widely used because an important goal of experimental neuroscience is to make inferences about channel tuning functions from neuroimaging data. The obvious way to do this would be to estimate the parameters of the tuning functions via model fitting to the data (e.g., using adaptive basis functions), and then make these the target of inference rather than the population responses. A problem with this solution is that encoding models are already complex, so adding free parameters is likely to increase the identifiability problems that already exist. Sadil, Huber, and Cowell (2021) recently addressed this issue by constraining the one-dimensional encoding model [see Equation (10.4)] in multiple ways. First, they assumed that tuning functions for all channels are identical except for their

preferred stimulus (i.e., homogeneous population code). Second, they avoided the many free weight parameters that characterize standard encoding models [as in the Equation (10.13) model] by assuming that the weights in each voxel follow a Gaussian-like curve centered at the stimulus value (e.g., orientation) that is preferred by the dominant channel in that voxel. Third, they limited the number of ways that the model predictions could be modified by some psychological or experimental factor (e.g., reducing stimulus contrast). In addition, they adopted a Bayesian framework that allowed them to introduce inferential biases through their chosen prior.

Inverted encoding modeling also falls victim to the feature fallacy error (Diedrichsen, 2020; Diedrichsen & Kriegeskorte, 2017). As explained earlier, an infinite number of channel response matrices can be chosen that produce exactly the same fit to the data (Gardner & Liu, 2019). Although these different channel responses all predict the same aggregate activity [see Equations (10.28)–(10.31)], their population response profiles can have dramatically different shapes. This highlights the fact that inverted encoding is only useful when the obtained estimates of the population responses are interpreted with specific reference to the tuning functions and other features of the model that was inverted (Sprague, Boynton, & Serences, 2019).

10.3.2 Stimulus Decoding and Reconstruction

The most common application of model inversion is not to estimate population responses, but either to decode stimulus values or to provide a full reconstruction of the presented stimulus. For example, the Equation (10.37) decoding scheme is easily extended to stimulus decoding – that is, from the problem of estimating the expected population response matrix E[$\mathbf{R}$] to the problem of testing the ability of the model to identify the stimuli that were presented during the test phase. The rows of the $\widehat{\mathbf{R}}_{\text{test}}$ matrix that results from applying Equation (10.37) will not exactly equal any of the rows of the expected channel-response matrix E[$\mathbf{R}_{train}$] that we constructed when building the Equation (10.13) encoding model (e.g., because of noise). So to use Equation (10.37) to complete the loop back to the stimulus, we need a classification scheme that will assign a single stimulus to each row of $\widehat{\mathbf{R}}_{\text{test}}$. Under the assumption that the noise vector $\underline{\boldsymbol{\epsilon}}_m$ in Equation (10.12) has a multivariate normal distribution in every voxel with mean vector $\underline{\mathbf{0}}$ and variance–covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathrm{I}$, it turns out that for each row in $\widehat{\mathbf{R}}_{\text{test}}$, the optimal classification strategy is to compute the correlation with every row in E[$\mathbf{R}_{train}$] and then associate that row in $\widehat{\mathbf{R}}_{\text{test}}$ with the row in E[$\mathbf{R}_{train}$] where the correlation is highest [assuming that the prior probabilities $P(\mathrm{S}_i)$ are equal for all stimuli; e.g., Fukunaga, 2013]. Row i of E[$\mathbf{R}_{train}$] contains the expected response of each channel to the presentation of stimulus S_i. Therefore, we can denote this row by $\mathbf{r}_{\text{train}}(\mathrm{S}_i)^\top$ [i.e., see Equation (10.12)]. Row m of $\widehat{\mathbf{R}}_{test}$ was generated by the mth event, but of course, we do not know which stimulus caused this event. So denote row m of $\widehat{\mathbf{R}}_{\text{test}}$ by $\widehat{\mathbf{r}}_{\text{test}}(\mathrm{E}_m)^\top$. Then the optimal decoding scheme uses the following classification rule.

Classify the mth event of the testing data as a stimulus S_i event if

$$\text{corr}\left[\widehat{\mathbf{r}}_{\text{test}}(E_m), \mathbf{r}_{\text{train}}(S_i)\right] = \max_{j=1,N_s} \text{corr}\left[\widehat{\mathbf{r}}_{\text{test}}(E_m), \mathbf{r}_{\text{train}}(S_j)\right]. \tag{10.38}$$

As indicated earlier, an encoding model is not necessary to perform stimulus decoding from fMRI data. This can also be achieved by training a machine-learning algorithm to extract information about stimuli from activity patterns. This type of nonparametric decoding appears in the literature more frequently than decoding by inverting an encoding model, but it has been argued that machine-learning approaches provide more limited opportunities to make inferences about underlying computational mechanisms (Kriegeskorte & Douglas, 2019; Naselaris *et al.*, 2011). In other words, a common assumption in the field is that although nonparametric decoding analyses can reveal *what* information is encoded in a given brain region, they cannot reveal information about *how* that information is encoded. On the other hand, experimental and modeling work reveals this to be at least partially incorrect.

For example, an important question in sensory neuroscience is whether a neural population encodes a stimulus property in a way that is invariant to some irrelevant stimulus change; that is, with encoding being the same across changes in an irrelevant feature. The opposite of such invariant encoding would be context-specific or configural encoding, in which the way a stimulus property is encoded by a population depends on the value of a second property. Both invariant and configural representations are important for discussions of how the brain represents objects and generalizes knowledge about them. Cognitive neuroscientists have used a variation of decoding analyses, called cross-decoding (or cross-classification, see Allefeld & Haynes, 2014; Anzellotti & Caramazza, 2014; Kaplan, Man, & Greening, 2015), to attempt to make inferences about invariant encoding in particular brain regions. The first step in cross-decoding is to train a classifier to decode a particular stimulus feature, such as the shape of an object, from patterns of fMRI activity observed across voxels. The second step is to test the trained classifier with new patterns of fMRI activity, this time obtained from presentation of the same stimuli, but changed in an irrelevant property, such as rotation in depth.

Theoretical and modeling work has shown that cross-decoding can indeed be used to make valid inferences about *how* stimuli are encoded in a particular area from neuroimaging data, without making any assumptions about specific aspects of the encoding model (Soto, Vucovich, & Ashby, 2018). However, cross-decoding provides evidence *against* the null hypothesis of context-specific encoding (i.e., generalization of decoding performance shows that encoding is not completely context-specific), and not evidence *for* the alternative of invariance. In addition, the test is prone to false positives because the measurement model can increase invariance in the transformation from neural to voxel space. Testing the null hypothesis of invariance in addition to cross-decoding allows one to reach more precise and valid conclusions about the underlying representations. These theoretical insights have been verified through experimental and simulation work

(Soto & Narasiwodeyar, 2021). It is likely that other general features of encoding can be inferred using nonparametric decoding, but more research is needed in this area.

In addition to simple decoding of the identity of a stimulus, model inversion can also be used for full stimulus reconstruction, thereby providing a method to visualize what has been encoded in the brain on a given trial. For example, Naselaris *et al.* (2009) used the structural model illustrated in Figure 10.1 and a Bayesian framework to reconstruct an image with the maximum posterior probability of having produced the measured BOLD activity. Their Bayesian framework allowed them to compare reconstruction under a variety of prior distributions over the images [$P(\mathrm{S}_i)$ in Equation (10.32)]. They found that reconstruction with a flat prior, which uses only information from voxel activities captured in the encoding model, was insufficient to reveal the identity of objects in the reconstructed images. A more informative prior that included some well-known statistical information about natural images (a $1/f$ amplitude spectrum and sparsity in the Gabor-wavelet domain) produced more natural-looking images, but still was unable to provide information about object identity. Finally, they attempted to better capture the prior distribution over natural images by sampling from it: they used a database of six million images as a prior, so that each image in the set had a prior probability of $\left(6 \times 10^6\right)^{-1}$, and any image outside this set had a prior probability of zero. This prior enabled them to reconstruct both the spatial structure and semantic content of the original images. A similar approach was used to reconstruct videos presented to participants from fMRI data (Nishimoto *et al.*, 2011).

More recent research in this area leverages the power of deep learning for image reconstruction, achieving reconstructions that could be recognized by humans without the need to sample explicitly from some pool of natural images (e.g., Ren *et al.*, 2021; Seeliger *et al.*, 2018; Shen *et al.*, 2019).

10.4 Representational Similarity Analysis

Representational similarity analysis (RSA) is a multivariate method that extracts similarity structures from BOLD activity (Kriegeskorte, Mur, & Bandettini, 2008). It identifies activation patterns that are similar and others that are dissimilar. A fundamental assumption is that two data sets that exhibit a comparable similarity structure must share a deeper homology in how the systems that generated those data represent and process events in the world. Perhaps the greatest strength of RSA is that a common approach can be used to extract similarity structures from many different modalities, allowing links to be drawn between vastly different levels of analysis. For example, consider a mathematical model of some perceptual or cognitive task that makes no neuroscience predictions per se, but instead assumes that performance depends on some hypothetical intervening variable, such as working memory load, attention, or reward prediction error. Next, suppose that for each pair of possible trial types, we use the model

to compute a predicted similarity by comparing its predicted values on the intervening variable on the two types of trials. We can then compare these predicted similarities to the similarity structure that RSA extracts from the BOLD data. If the similarities predicted by the model and the similarity structure derived from the BOLD responses in some ROI are qualitatively similar, then RSA concludes that this ROI may play a key role in computing the value of the hypothesized intervening variable.

RSA is conceptually simple. The first step is to compute a representational dissimilarity matrix (RDM), which includes a row and column for every event, condition, ROI, or task, depending on what type of similarity structure we want to construct. For the present purposes, there are three obvious possibilities. One is that the RDM will include dissimilarities between all possible pairs of activity patterns estimated from a voxel-based encoding model (i.e., rows of $\widehat{\mathbf{A}}$). Another possibility is that the RDM is estimated directly from the BOLD data in some ROI for the same events that were used to create the activity patterns. Finally, a third possibility is that the RDM is constructed from some other type of mathematical model – for example, a traditional model of perceptual or cognitive processing from the mathematical psychology literature. However the RDM is created, it is assumed to include numerical data that define the similarity structure describing how the various events are related.

The RDM is sometimes used to build a similarity structure using some form of multidimensional scaling. But in most applications, two RDMs of the same task are directly compared. For example, RSA is often used to test the validity of an encoding model by testing statistically whether the RDM predicted by the model is consistent with an empirical RDM estimated in some ROI from our fMRI data.

10.4.1 Estimating an RDM

An RDM is estimated by computing the dissimilarity in the model predictions or data for all possible pairs of stimulus types (or more generally, event types). If there are N_S different stimuli, then these dissimilarities are collected in an RDM of order $N_S \times N_S$. The entry in row i and column j is the observed (in the case of BOLD data) or predicted (in the case of a model) dissimilarity between the response to stimulus types i and j. Denote this dissimilarity by $d(\mathrm{S}_i, \mathrm{S}_j)$.

In the case of BOLD data, $d(\mathrm{S}_i, \mathrm{S}_j)$ is computed by comparing rows of the BOLD activity matrix $\tilde{\mathbf{B}}$. Recall that $\tilde{\mathbf{B}}$ is an $N_S \times N_v$ matrix in which row i and column k contains the estimated amplitude of the BOLD response to stimulus S_i in voxel k of the ROI. Therefore, row i is a vector describing the response of the ROI to stimulus S_i. In the case of voxel-based encoding models, the predicted aggregate activity vector $\mathbf{A}$ replaces $\tilde{\mathbf{B}}$. In the case of more traditional mathematical psychology models, the RDM is computed by comparing predictions of the model – usually on the intervening variable of interest – on all possible pairs of stimulus trials.

Let $\underline{\mathbf{a}}_i^\top$ denote the ith row of the aggregate activity matrix $\mathbf{A}$ predicted by some voxel-based encoding model. Then $d(\mathrm{S}_i, \mathrm{S}_j)$ is an estimate of the dissimilarity of $\underline{\mathbf{a}}_i$

and $\underline{\mathbf{a}}_j$. The concept of similarity is fundamentally important in almost every scientific field. And across these different fields, similarity and dissimilarity are defined in many different ways. In RSA, the choice of the best dissimilarity measure is still an area of active research (Bobadilla-Suarez *et al.*, 2020). Most applications, however, have used one of three different measures: one minus the Pearson correlation, a Euclidean-distance measure, or a Mahalanobis-distance measure.

As the name suggests, one minus the Pearson correlation equals

$$d_{\mathrm{P}}(\mathrm{S}_i, \mathrm{S}_j) = 1 - r(\underline{\mathbf{a}}_i, \underline{\mathbf{a}}_j), \tag{10.39}$$

where $r(\underline{\mathbf{a}}_i, \underline{\mathbf{a}}_j)$ is the Pearson correlation between the entries in $\underline{\mathbf{a}}_i$ and $\underline{\mathbf{a}}_j$. The Euclidean measure is defined as the squared Euclidean distance between $\underline{\mathbf{a}}_i$ and $\underline{\mathbf{a}}_j$:

$$d_{\mathrm{E}}(\mathrm{S}_i, \mathrm{S}_j) = (\underline{\mathbf{a}}_i - \underline{\mathbf{a}}_j)^\top (\underline{\mathbf{a}}_i - \underline{\mathbf{a}}_j). \tag{10.40}$$

Mahalanobis dissimilarity is based on the assumption that the underlying data are samples from a multivariate normal distribution. The Mahalanobis dissimilarity between activity vectors $\underline{\mathbf{a}}_i$ and $\underline{\mathbf{a}}_j$, which is defined as the squared Mahalanobis distance between the vectors, equals

$$d_{\mathrm{M}}(\mathrm{S}_i, \mathrm{S}_j) = (\underline{\mathbf{a}}_i - \underline{\mathbf{a}}_j)^\top \widehat{\mathbf{\Sigma}}^{-1} (\underline{\mathbf{a}}_i - \underline{\mathbf{a}}_j), \tag{10.41}$$

where $\widehat{\Sigma}^{-1}$ is the inverse of the estimated (spatial) variance–covariance matrix of the activity vectors.

One weakness of all these measures is that if two activity vectors are identical at the population level, and therefore their distance apart is zero, then noise can only increase the distance between them. Therefore, under the null hypothesis that two event types elicit identical activity patterns, all of these distance measures will produce biased estimates of the true difference. One solution to this problem is to use cross-validated Mahalanobis distance, or crossnobis distance (Allefeld & Haynes, 2014). The crossnobis distance is computed by dividing the data into Q independent partitions, and using a leave-one-partition-out scheme. Let $\underline{\mathbf{a}}_i(q)$ denote the ith activity pattern computed from the data in partition q, and $\underline{\mathbf{a}}_i(\neg q)$ denote the same activity computed from the data in all partitions other than q. Then the cross-validated Mahalanobis distance – that is, the crossnobis distance – between the activity vectors associated with stimuli S_i and S_i is

$$d_{\mathrm{CN}}(\mathrm{S}_i, \mathrm{S}_j) = \frac{1}{Q} \sum_{q=1}^{Q} [\underline{\mathbf{a}}_i(q) - \underline{\mathbf{a}}_j(q)]^\top \widehat{\mathbf{\Sigma}}^{-1} [\underline{\mathbf{a}}_i(\neg q) - \underline{\mathbf{a}}_j(\neg q)]. \tag{10.42}$$

Note that because $[\underline{\mathbf{a}}_i(k) - \underline{\mathbf{a}}_j(k)]$ and $[\underline{\mathbf{a}}_i(\neg k) - \underline{\mathbf{a}}_j(\neg k)]$ are computed from different data partitions, the crossnobis distance $d_{\mathrm{CN}}(\mathrm{S}_i, \mathrm{S}_j)$ could be either positive or negative. In contrast, of course, regular Euclidean and Mahalanobis distance must both always be non-negative. The advantage of crossnobis distance is that it eliminates bias. More specifically, under the null hypothesis that two events elicit the same pattern of activation, the mean crossnobis distance between the resulting activity vectors is zero, whereas with regular Euclidean or Mahalanobis distance,

this mean is greater than zero (Allefeld & Haynes, 2014). Furthermore, A. Walther *et al.* (2016) compared all of these measures on simulated and real fMRI data. The most reliable method was crossnobis distance. Even so, the choice of the best dissimilarity measure is still an area of active research. While crossnobis distance has the appealing property of being unbiased and has been shown to be more reliable than other measures, some researchers have recently argued that the one-minus-Pearson-correlation measure is preferable (Bobadilla-Suarez *et al.*, 2020).

10.5 Testing Encoding Models Against Behavioral Data

The introduction to this chapter claimed that many of the identifiability problems that plague computational models of behavior could be alleviated by extending tests of the models to fMRI data. However, we also saw that encoding models have their own identifiability problems that complicate their interpretation. Even so, it now seems clear that an integrative approach, in which behavioral and neuroimaging data are both addressed within the same modeling framework, would be beneficial in both mathematical psychology and computational neuroimaging (Soto, 2019).

There are at least three ways in which encoding models can be tested against behavioral data. First, we can use encoding models that are grounded in neuroscience to predict behavioral data. Second, we can fit a cognitive model to behavioral data, build an encoding model in which the encoding channels compute the intervening variables hypothesized by the cognitive model, and then test the resulting encoding model against fMRI data. This approach is known as model-based fMRI (O'Doherty, Hampton, & Kim, 2007). Third, we can jointly model fMRI and behavioral data in a truly integrative approach that constrains inferences about a single model with both types of data. We now briefly describe each of these approaches.

10.5.1 Encoding/Decoding Observer Models

One way to build an encoding model that makes simultaneous neural and behavioral predictions is to generalize any of the voxel-based encoding models described earlier in a way that allows them to make behavioral predictions. In all of those models, the population neural response vector $\mathbf{r}$ is assumed to be available to downstream neurons to decode useful behavioral information about the stimulus. So to make behavioral predictions, two additional problems must be solved. First, a choice must be made about which of a variety of possible decoding schemes is incorporated into the model (e.g., Lehky, Sereno, & Sereno, 2013; Pouget *et al.*, 1998; Salinas & Abbott, 1994; Seung & Sompolinsky, 1993). Second, assumptions must be made about how the model uses the decoded stimulus information to select a response. We refer to encoding models that add a decoding scheme and response selection assumptions as *encoding/decoding observer models*.

As an illustration of this approach, consider a simple identification task in which the stimuli vary on a single physical dimension [e.g., as in Equation (10.4)]. For example, the stimuli might all be Gabor patterns that vary only on orientation or spatial frequency. The question of which decoding scheme to use is complicated somewhat by the fact that some schemes lead to an inherent ambiguity in whether an observed behavioral change is due to encoding versus decoding changes (Gold & Ding, 2013). Confronted with this dilemma, many modelers have assumed optimal decoding via maximum likelihood estimation (e.g., Dakin, Mareschal, & Bex, 2005; Deneve, Latham, & Pouget, 1999; Hays & Soto, 2020; Ling, Liu, & Carrasco, 2009; May & Solomon, 2015; Paradiso, 1988; Series, Stocker, & Simoncelli, 2009; Soto *et al.*, 2021). This assumption leads to the decoding scheme in which observation of the neural response vector $\mathbf{r}$ causes the model to infer that the value of the presented stimulus was $\hat{s}$, where

$$\hat{s} = \arg\max_s \hat{P}(s|\mathbf{r}, \underline{\boldsymbol{\theta}}), \tag{10.43}$$

and as usual, $\underline{\boldsymbol{\theta}}$ is a vector of channel tuning parameters.

If neural noise is independent across channels, then

$$P(s|\mathbf{r}, \underline{\boldsymbol{\theta}}) = \prod_{c=1}^{N_c} P(s|r_c, \underline{\boldsymbol{\theta}}), \tag{10.44}$$

and therefore, the log-likelihood is maximized when

$$\hat{s} = \arg\max_s \sum_{c=1}^{N_c} \ln \hat{P}(s|r_c, \underline{\boldsymbol{\theta}}). \tag{10.45}$$

There is usually a single optimal solution for a well-posed statistical problem such as this, which therefore avoids the ambiguities mentioned above about whether behavioral changes are caused by encoding or decoding mechanisms. An additional advantage is that the asymptotic properties of maximum likelihood estimators are well known. In particular, maximum likelihood estimators are asymptotically normal, and if noise is independent and identically distributed across channels, then the maximum likelihood estimator $\hat{s}$ of the true stimulus value s_0 has an asymptotic normal distribution with mean s_0 and variance

$$\sigma_{\hat{s}}^2 = [n\, I(s_0)]^{-1},$$

where $I(s_0)$ is the Fisher information, and n is the number of channels (e.g., Van der Vaart, 2000).

Note that this variance can be directly computed if an analytical form for the Fisher information is known, which is the case for the standard encoding model with Gaussian tuning functions that all have identical width ω [i.e., see Equation (10.4)]. When, in addition, neural noise is Poisson and independent, the Fisher information is given by (Dayan & Abbott, 2001; Pouget *et al.*, 1998; Seung & Sompolinsky, 1993):

$$I(s) = \sum_{c=1}^{N} \frac{\left[f_c'(s)\right]^2}{f_c(s)}$$
$$= \sum_{c=1}^{N} \frac{r^{max}(s - s_c)^2}{\omega^4} \exp\left[-\frac{1}{2}\left(\frac{s - s_c}{\omega}\right)^2\right], \tag{10.46}$$

where $f_c(s)$ is the Gaussian tuning function of Equation (10.4) and $f_c'(s)$ is its derivative with respect to s. For Gaussian neural noise with fixed variance σ_r^2, the Fisher information is given by (Pouget *et al.*, 1998):

$$I(s) = \frac{1}{\sigma_r^2} \sum_{c=1}^{N} f_c'(s)^2$$
$$= \frac{1}{\sigma_r^2} \sum_{c=1}^{N} \frac{r^{max}(s - s_c)^2}{\omega^4} \exp\left[-\left(\frac{s - s_c}{\omega}\right)^2\right]. \tag{10.47}$$

When $I(s)$ is unknown, which is likely to be the case for many encoding models, $\sigma_{\hat{s}}^2$ can be directly estimated through Monte Carlo simulation (e.g., Dakin, Mareschal, & Bex, 2005; Hays & Soto, 2020; Ling, Liu, & Carrasco, 2009).

Another advantage of assuming that decoding is optimal is that it allows encoding/decoding observer models to be linked to psychophysical measures in a straightforward manner. For example, the distribution of $\hat{s}$ could be interpreted as the distribution of perceptual evidence assumed by Gaussian signal detection theory (Ashby & Wenger, Chapter 6 of this volume; Green & Swets, 1966; Macmillan & Creelman, 2005), which links the encoding/decoding observer model to popular measures such as d' and sensory thresholds. For example, consider a two-stimulus identification task with stimuli that have values s_1 and s_2. Suppose we use these asymptotic results to compute the mean $\mu_{\hat{s}}$ and variance $\sigma_{\hat{s}}^2$ of the distribution of estimates for each stimulus, either through analytical expressions or Monte Carlo simulation. From these values, it is easy to compute the model's predicted d' for the identification task (Soto *et al.*, 2021):

$$d' = \frac{\mu_{\hat{s}_1} - \mu_{\hat{s}_2}}{\sqrt{0.5\left(\sigma_{\hat{s}_1}^2 + \sigma_{\hat{s}_2}^2\right)}}.$$

Note that $I(s)$ is a function of the stimulus value, so the variance of decoded values might change when different stimuli are presented. However, most researchers assume that it remains the same across values of the decoded variable, in line with the equal-variance signal detection model.

The methods used to create encoding/decoding observer models allow behavioral predictions to be generated from almost any encoding model that has either been fitted to neural data or constrained by it. For example, Goris *et al.* (2013) showed that an encoding/decoding observer model constrained by what is known about encoding of spatial frequency in primary visual cortex does an excellent job

at predicting pattern detection behavior. In principle, any well-defined encoding model can serve as a model of behavior with relatively minor adjustments.

Equations (10.28)–(10.31) showed that many different sets of encoding channels make identical predictions. This can make it difficult to draw strong inferences about why some change occurred in a population response. One way to resolve these ambiguities is to explore various alternatives by formulating them as hypotheses that make distinct behavioral predictions in some psychophysical task. Simulation work has shown that when combined with inverted encoding modeling, only a couple of psychophysical experiments are sufficient to arbitrate between major hypotheses about changes in neural encoding (Hays & Soto, 2020).

Signal detection theory has been an invaluable model, not only in perceptual tasks, but also in cognitive tasks such as recognition memory (e.g., Wixted, 2007), causal and contingency learning (e.g., Siegel *et al.*, 2009), generalization (e.g., Blough, 1967), and metacognition (e.g., Maniscalco & Lau, 2012, 2014). For this reason, the methods that have been successfully used to link encoding models to psychophysics in the vision literature might prove useful in other research areas as well.

10.5.2 Model-Based FMRI

All of the encoding models considered so far were designed specifically with the goal of modeling fMRI data. But fMRI data can also be used to provide unique tests of cognitive-based mathematical models that are more traditional within mathematical psychology. The methods that have been developed to test the validity of purely behavioral computational models against fMRI data are known as *model-based fMRI* (O'Doherty, Hampton, & Kim, 2007).

Purely behavioral models are those that make no neuroscience predictions. Instead, they typically make predictions about how a participant will respond to a stimulus by appealing to some hypothetical constructs or latent (intervening) variables, such as, for example, memory, attention, or similarity. The models are tested against behavioral data by examining their ability to account for dependent variables such as response accuracy and response time. A good fit provides only indirect support for the model and its hypothesized latent variables – in part, because of the identifiability problems described earlier. Model-based fMRI provides an opportunity to improve model identifiability by offering a method to examine the latent variables more directly. The basic idea is to estimate the free parameters of the model by fitting it to the available behavioral data – in exactly the same way that the model is typically applied. Next, the parameter estimates that result are used to derive predictions from the model about one or more latent variables, and finally these predictions are compared to the observed BOLD responses from various brain regions (e.g., by using the GLM). For example, consider an exemplar model that predicts trial-by-trial categorization responses are determined by certain specific similarity computations. In model-based fMRI, the critical similarity value predicted by the model is computed on every trial and then

correlated with trial-by-trial observed BOLD responses, either across the whole brain or in specific brain regions. Finding a region where the correlation is high accomplishes two goals. First, it provides empirical support for the model that is impossible with purely behavioral data because it suggests that changes in neural activity in some brain region are consistent with changes in a latent variable that the model predicts is critical to the task under study. Second, a good fit identifies brain regions that might possibly mediate the processes hypothesized by the model. Since the models are perceptual or cognitive, this allows an important first step in extending them to the neural level.

In the ideal application, the constructs that are tested against fMRI data change significantly from trial to trial. For example, consider a model that assumes participants compare the presented stimulus to some internally constructed decision criterion and give one response if the criterion is exceeded and a different response if it is not (e.g., as in signal detection theory). A model that predicts the numerical value of this criterion on every trial could be tested against fMRI data by correlating the predicted criterion value against the BOLD response observed in different brain regions. However, if the experimental design is such that the model predicts only slow changes in the criterion during the scanning session, then these correlations will not provide strong tests of the model because the predicted BOLD responses in criterion-setting regions will be similar to the BOLD responses in task-inactive brain regions.

After some model-predicted hypothetical constructs are selected that vary significantly from trial to trial, a typical model-based fMRI analysis would include the following steps. First, the model is fit to the behavioral data collected during the functional run separately for each participant. The primary purpose of this step is to estimate the free parameters in the model. Since the model being tested is purely behavioral, it makes no predictions about neural activations or BOLD responses, and as a result, its parameters should only be estimated by fitting to behavioral data.

The second step is to use the parameter estimates from step one to compute numerical values of the intervening variables from the model that were identified earlier to test against the fMRI data. The goal here is to identify brain regions in which changes in the BOLD responses are predicted by changes in the variables. In the case of the exemplar model, obvious candidates include the predicted summed similarity of the presented stimulus to each of the contrasting categories.

Step three is to construct a model of the BOLD response from each of the selected model variables. The standard approach is to first construct a boxcar function of square waves for each variable. The height of this function is set to zero when the variable is predicted to be inactive and to the value of the variable when it is active. For example, in the case of the exemplar model's predicted summed similarity to some category A, the boxcar function would equal zero between trials and its height would equal the predicted summed similarity to exemplars from category A during the time beginning with each stimulus onset and ending with the participant's response. After this boxcar function is built, predicted BOLD

responses are computed by convolving the boxcar function with some model of the hrf [as in Equation (10.21)].

Step four is to correlate each of these predicted BOLD responses with the observed BOLD response in every voxel via the GLM. Voxels where the correlation is high are identified as being sensitive to that variable (for a more thorough description of all these steps see, e.g., Ashby, 2019).

In summary, a model-based fMRI analysis of this type: (1) tests the model against a new dependent variable (i.e., the BOLD response); (2) potentially makes the model's latent variables observable; (3) identifies brain regions sensitive to the model's latent variables; and (4) provides valuable data that could be used to develop a neurocomputational version of the model.

10.5.3 Joint Neural and Behavioral Modeling

Encoding/decoding observer models are neural models in which some assumptions are added that allow tests against behavioral data. In contrast, model-based fMRI is an approach in which assumptions are added to purely behavioral models that allow tests against fMRI data. A third way in which encoding models can be tested against behavioral data is to build models that directly account for both neuroscience and behavioral data. There are two general approaches to joint modeling of this kind – one based in neuroscience and one based in statistics. Their main advantage is that they use variation in both behavioral and neural data to jointly and equally constrain inferences about encoding models.

The neuroscience approach comes from the emerging field of computational cognitive neuroscience (CCN), which is a new field that lies at the intersection of computational neuroscience, machine learning, and neural network theory (i.e., connectionism) (Ashby, 2018; O'Reilly & Munakata, 2000). The goal here is to build biologically detailed neural network models in which the simulated regions and their interconnections are faithful to known neuroanatomy. The units that define the network are either simulated spiking neurons or populations of similar neurons (e.g., a cortical column), in which case the primary dependent variables are the firing rates of each population. Theoretically at least, CCN models can account for all levels of a behavioral phenomenon from single-neuron spiking up to behavior. In particular, a good CCN model should predict how neural activity changes in a variety of different brain regions as the subject performs the task under study, and at the same time make predictions about the most widely studied behavioral dependent variables, including response accuracy and response time. In general, testing CCN models against fMRI data follows the same basic steps as in model-based fMRI. For a description of the special issues that arise due to the extra neuroscience details of CCN models, see Ashby (2019).

The statistical approach to joint modeling uses a hierarchical Bayesian inferential framework to model the statistical relations between neural and behavioral measures directly within a single model (Palestro *et al.*, 2018; Turner, 2015; Turner *et al.*, 2013). To keep the presentation concrete, consider an identification

experiment in which participants are presented with one of two stimuli on each trial, S_1 and S_2, and their task is to report which of the two stimuli was presented. Model performance in this task will depend on the specific stimuli that are presented and their base rates, which can be collected in the set $\mathcal{S} = \{S_1, S_2, P(S_1), P(S_2)\}$. The neural dependent variables are the amplitudes of the BOLD responses to the two stimuli, collected in the 2×1 vector $\underline{\tilde{\mathbf{b}}}$, and the behavioral dependent variables are the proportion of correct responses on S_1 trials and on S_2 trials, which can be collected in a 2×1 vector $\underline{\mathbf{o}}$. Finally, we assume that the BOLD responses are related to the channel responses according to the linearized encoding model of Equation (10.15).

To build a joint model, we begin by computing the likelihood of the fMRI data, $P(\underline{\tilde{\mathbf{b}}}|\mathbf{R}, \underline{\boldsymbol{\beta}}, \mathcal{S})$, where $\underline{\boldsymbol{\beta}}$ represents a vector of parameters from the neural measurement model. For example, in the linearized encoding model, $\underline{\boldsymbol{\beta}}$ would include the weight parameters in $\underline{\mathbf{w}}$ as well as the variance–covariance matrix of measurement noise $\boldsymbol{\Sigma}_m$. Second, we compute the likelihood of the behavioral data, $P\left(\underline{\mathbf{o}}|\mathbf{R}, \underline{\boldsymbol{\gamma}}, \mathcal{S}\right)$, where $\underline{\boldsymbol{\gamma}}$ is a vector of parameters from the behavioral measurement model. Both of these likelihoods depend directly on the random population response $\mathbf{R}$, which has a distribution $P\left(\mathbf{R}|\underline{\boldsymbol{\theta}}, \mathcal{S}\right)$ specified either by Equation (10.2) or (10.3), and that depends on the encoding model parameters and the stimulus set $\mathcal{S}$ (we omit state variables for simplicity). Finally, the model should formalize prior distributions over all the parameters included in $\underline{\boldsymbol{\theta}}$, $\underline{\boldsymbol{\beta}}$, and $\underline{\boldsymbol{\gamma}}$, which would depend on hyperparameters $\boldsymbol{\Omega}$. With this, the model is fully specified and the joint posterior distribution of the model parameters can be expressed as

$$P\left(\underline{\boldsymbol{\theta}}, \underline{\beta}, \underline{\boldsymbol{\gamma}} \mid \underline{\tilde{\mathbf{b}}}, \underline{\mathbf{o}}\right) \propto P\left(\underline{\tilde{\mathbf{b}}} \mid \mathbf{R}, \underline{\beta}, \mathcal{S}\right) P\left(\underline{\mathbf{o}} \mid \mathbf{R}, \underline{\boldsymbol{\gamma}}, \mathcal{S}\right) P\left(\mathbf{R} \mid \mathcal{S}, \underline{\theta}\right) P\left(\underline{\boldsymbol{\theta}}, \underline{\beta}, \underline{\boldsymbol{\gamma}} \mid \boldsymbol{\Omega}\right). \tag{10.48}$$

In general, this distribution can be approximated using any of a wide range of available sampling algorithms (see Gilks *et al.*, 1996).

Under the assumption that the BOLD responses are related to the channel responses according to the linearized encoding model of Equation (10.15), the likelihood of the BOLD amplitude $P\left(\tilde{\mathbf{b}}_t|\mathbf{R}, \underline{\beta}, \mathcal{S}\right)$ is multivariate Gaussian with mean $\mathrm{E}[\mathbf{R}]\ \underline{\mathbf{w}}$ [i.e., see Equation (10.15)] and variance–covariance matrix $\boldsymbol{\Sigma}_m$. The priors over $\underline{\mathbf{w}}$ and $\boldsymbol{\Sigma}_m$ can be chosen to match the regularization algorithms used in past applications of encoding modeling (Diedrichsen & Kriegeskorte, 2017), or to be conjugate for the likelihood function, which facilitates inference. The likelihood of the behavioral data $P\left(\underline{\mathbf{o}} \mid \mathbf{R}, \underline{\boldsymbol{\gamma}}, \mathcal{S}\right)$ can be obtained by linking the encoding model to signal detection theory in the way described earlier in this section. In this approach, an optimal decoder is used to obtain estimates of the noise in the decoded stimuli. With the addition of a threshold parameter, one can obtain the likelihood of each possible response on a given trial from the cumulative normal distribution. As before, priors can be chosen following previous applications of signal detection theory that have used a Bayesian framework, or to be conjugate to the likelihood function. Finally, the distribution of population responses $P\left(\mathbf{R} \mid \mathcal{S}, \underline{\theta}\right)$ will depend on our choice of tuning functions and neural noise, and priors can be chosen to be

conjugate to that distribution, or based on previous applications (Sadil, Huber, & Cowell, 2021; Van Bergen *et al.*, 2015).

10.6 Conclusions

Mathematical psychologists build and test mathematical models of perceptual, cognitive, and motor behaviors. A common goal is to develop models that describe the underlying processes that are presumed to mediate the behavior under study. When tested in the traditional way – that is, against behavioral measures such as response accuracy and response time – these processes are almost always unobservable. One common barrier that limits progress in this field is that models postulating very different psychological processes can often provide a similarly good quantitative fit to the behavioral data. For example, because of such nonidentifiabilities, many subfields are still debating the validity of competing models that were proposed 40 and 50 years ago.

Testing these models against fMRI BOLD data offers the hope of greatly improving model identifiability. And, because of methods such as RSA and model-based fMRI, this is true even for models that include no neuroscience detail. For example, any model that makes predictions about psychological processes that are unobservable with behavioral data could benefit from testing via model-based fMRI, at least so long as those predictions change significantly trial-by-trial. In particular, if two competing models account for behavioral data about equally well, then we should favor the model that makes predictions about trial-by-trial changes in some psychological process that track changes in the BOLD response of some brain region, over the model that makes process predictions that are not mirrored by BOLD data.

As an example of how RSA might benefit cognitive modeling, suppose some cognitive theory predicts that the same perceptual and cognitive processes mediate performance in two different tasks. Then this theory should predict similar patterns of activation in an fMRI study of the two tasks, even if the theory makes no predictions about what those activation patterns should look like. If an RSA concludes that the activation patterns in the two tasks are qualitatively different, then the theory probably needs some significant revision.

Although the number likely decreases every year, there are still many cognitive scientists who are deeply skeptical of fMRI – some even characterizing it as a new form of phrenology (Dobbs, 2005; Uttal, 2001). Even so, recent methodological advancements, such as model-based fMRI and RSA, show that fMRI can provide useful and powerful new tests of models – even purely cognitive models – that would have been considered a fantasy just a few decades ago.

10.7 Related Literature

For a thorough description of virtually all statistical methods for analyzing fMRI BOLD data – including traditional GLM approaches, as well as encoding and decoding methods, RSA, and DCM – see Ashby (2019).

An introduction to encoding and decoding from a computational neuroscience perspective can be found in Pouget, Dayan, and Zemel (2003) and Dayan and Abbott (2001). For an introduction to applications of encoding models to neuroimaging, see van Gerven (2017).

Decoding analyses of neuroimaging data using machine-learning algorithms (e.g., MVPA) rather than explicit encoding modeling are covered by Pereira, Mitchell, and Botvinick (2009). Kriegeskorte and Diedrichsen (2019) summarize recent work on RSA and its relation to encoding modeling (see also Diedrichsen & Kriegeskorte, 2017). May and Solomon (2015) describe encoding/decoding observer modeling in detail, and O'Doherty *et al.* (2007) does the same for model-based fMRI. Palestro *et al.* (2018) give a tutorial introduction to joint modeling of neural and behavioral data using a hierarchical Bayesian framework.

Acknowledgments

We thank Thomas Sprague, Justin Gardner, and Joshua Ryu for their helpful comments on an earlier version of this chapter.

References

Allefeld, C., & Haynes, J. D. (2014, April). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage*, *89*, 345–357. doi: 10.1016/j.neuroimage.2013.11.043

Anzellotti, S., & Caramazza, A. (2014). The neural mechanisms for the recognition of face identity in humans. *Frontiers in Psychology*, *5*, 672. doi: 10.3389/fpsyg.2014.00672

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F. G. (2018). Computational cognitive neuroscience. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology,* vol. 2 (pp. 223–270). New York: Cambridge University Press.

Ashby, F. G. (2019). *Statistical analysis of fMRI data*, 2nd ed. Cambridge, MA: MIT Press.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632–656.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*(4), 423–466.

Bickel, P. J., & Li, B. (2006). Regularization in statistics. *Test*, *15*(2), 271–344.

Blough, D. S. (1967). Stimulus generalization as signal detection in pigeons. *Science*, *158*(3803), 940–941.

Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., & Love, B. C. (2020). Measures of neural similarity. *Computational Brain & Behavior*, *3*, 369–383.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, *16*(13), 4207–4221.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience*, *29*(44), 13992–14003. doi: 10.1523/JNEUROSCI.3577-09.2009

Buhmann, M. D. (2003). *Radial basis functions: Theory and implementations* (Vol. 12). Cambridge, MA: Cambridge University Press.

Buxton, R. B. (2013). The physics of functional magnetic resonance imaging (fMRI). *Reports on Progress in Physics*, *76*(9), 096601.

Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magnetic Resonance in Medicine*, *39*(6), 855–864.

Byers, A., & Serences, J. T. (2014). Enhanced attentional gain as a mechanism for generalized perceptual learning in human visual cortex. *Journal of Neurophysiology*, *112*(5), 1217–1227. doi: 10.1152/jn.00353.2014

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. doi: 10.1038/nrn3136

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2013). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, *26*(1), 132–142.

Chen, C. T. (1970). *Introduction to linear systems theory*. New York: Holt, Rinehart & Winston.

Chen, Y., Geisler, W. S., & Seidemann, E. (2006). Optimal decoding of correlated neural population responses in the primate visual cortex. *Nature Neuroscience*, *9*(11), 1412–1420.

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770.

Dakin, S. C., Mareschal, I., & Bex, P. J. (2005). Local and global limitations on direction integration assessed using equivalent noise analysis. *Vision Research*, *45*(24), 3027–3049. doi: 10.1016/j.visres.2005.07.037

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

DeCharms, R. C. (2008). Applications of real-time fMRI. *Nature Reviews Neuroscience*, *9*(9), 720–729.

Deneve, S., Latham, P. E., & Pouget, A. (1999). Reading population codes: A neural implementation of ideal observers. *Nature Neuroscience*, *2*(8), 740–745. doi: 10.1038/11205

Diedrichsen, J. (2020). Representational models and the feature fallacy. In D. Poeppel, G. R. Mangun, & M. S. Gazzaniga (Eds.), *The cognitive neurosciences*, 6th ed. (pp. 669–678). Cambridge, MA: MIT Press.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, *13*(4), e1005508. doi: 10.1371/journal.pcbi.1005508

Dobbs, D. (2005). Fact or phrenology? *Scientific American Mind*, *16*(1), 24–31.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, *39*(2), 647–660.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. (2013). A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience*, *25*(5), 754–761.

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron*, *87*(4), 893–905.

Ester, E. F., Sprague, T. C., & Serences, J. T. (2020). Categorical biases in human occipitoparietal cortex. *Journal of Neuroscience*, *40*(4), 917–931.

Fracasso, A., Dumoulin, S. O., & Petridou, N. (2021). Point-spread function of the BOLD response across columns and cortical depth in human extra-striate cortex. *Progress in Neurobiology*, 102034.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*, 2nd ed. San Diego, CA: Academic Press.

Garcia, J. O., Srinivasan, R., & Serences, J. T. (2013). Near-real-time feature-selective modulations in human cortex. *Current Biology*, *23*(6), 515–522. doi: 10.1016/j.cub.2013.02.013

Gardner, J. L., & Liu, T. (2019). Inverted encoding models reconstruct an arbitrary model response, not the stimulus. *eNeuro*, *6*(2), e0363–18.2019. doi: 10.1523/ENEURO.0363-18.2019

Gilks, W. R., Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114. doi: 10.1016/j.pneurobio.2012.05.008

Goris, R. L. T., Putzeys, T., Wagemans, J., & Wichmann, F. A. (2013). A neural population model for visual pattern detection. *Psychological Review*, *120*(3), 472–496.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, *179*, 252–262. doi: 10.1016/j.neuroimage.2018.06.022

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534.

Hays, J. S., & Soto, F. A. (2020). Changes within neural population codes can be inferred from psychophysical threshold studies. *bioRxiv*, 2020.03.26.010900. doi: 10.1101/2020.03.26.010900

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Kaplan, J. T., Man, K., & Greening, S. G. (2015). Multivariate cross-classification: Applying machine learning techniques to characterize abstraction in neural representations. *Frontiers in Human Neuroscience*, *9*, 151. doi: 10.3389/fnhum .2015.00151

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual Review of Neuroscience*, *42*, 407–432.

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167–179. doi: 10.1016/j.conb.2019.04.002

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lee, S., Papanikolaou, A., Logothetis, N. K., Smirnakis, S. M., & Keliris, G. A. (2013). A new method for estimating population receptive field topography in visual cortex. *NeuroImage*, *81*, 144–157.

Lehky, S. R., Sereno, M. E., & Sereno, A. B. (2013). Population coding and the labeling problem: Extrinsic versus intrinsic representations. *Neural Computation*, *25*(9), 2235–2264.

Ling, S., Liu, T., & Carrasco, M. (2009). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Research*, *49*(10), 1194–1204.

Liu, T., Cable, D., & Gardner, J. L. (2018). Inverted encoding models of human population response conflate noise and neural tuning width. *Journal of Neuroscience*, *38*(2), 398–408.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*, 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d′, response-specific meta-d′, and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Berlin: Springer.

May, K. A., & Solomon, J. A. (2015). Connecting psychophysical performance to neuronal response properties I: Discrimination of suprathreshold stimuli. *Journal of Vision*, *15*(6), 8. doi: 10.1167/15.6.8

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643.

Murdock, B. B. (1985). An analysis of the strength-latency relationship. *Memory & Cognition*, *13*(6), 511–521.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

O'Doherty, J., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*(1), 35–53.

Ogawa, S., Lee, T.-M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, *87*(24), 9868–9872.

Ogawa, S., Lee, T.-M., Nayak, A. S., & Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, *14*(1), 68–78.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z. L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20–48. doi: 10.1016/j.jmp.2018.03.003

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics*, *58*(1), 35–49.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, *104*, 209–220. doi: 10.1016/j.neuroimage.2014.09.060

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1), S199–S209.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*(2), 125–132.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, *26*(1), 381–410.

Pouget, A., Zhang, K., Deneve, S., & Latham, P. E. (1998). Statistically efficient estimation using population coding. *Neural Computation*, *10*(2), 373–401.

Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., & Gao, X. (2021). Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, *228*, 117602. doi: 10.1016/j.neuroimage.2020.117602

Ritchie, J. B., & Carlson, T. A. (2016). Neural decoding and "inner" psychophysics: A distance-to-bound approach for linking mind, brain, and behavior. *Frontiers in Neuroscience*, *10*. doi: 10.3389/fnins.2016.00190

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology*, *11*(6), e1004316. doi: 10.1371/journal.pcbi.1004316

Sadil, P., Huber, D. E., & Cowell, R. A. (2021). NeuroModulation Modeling (NMM): Inferring the form of neuromodulation from fMRI tuning functions. *bioRxiv*, 2021.03.04.433362. doi: 10.1101/2021.03.04.433362

Salinas, E., & Abbott, L. F. (1994). Vector reconstruction from firing rates. *Journal of Computational Neuroscience*, *1*(1), 89–107.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. A. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785. doi: 10.1016/j.neuroimage.2018.07.043

Serences, J. T., Saproo, S., Scolari, M., Ho, T., & Muftuler, L. T. (2009). Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage*, *44*(1), 223–231. doi: 10.1016/j.neuroimage .2008.07.043

Series, P., Stocker, A. A., & Simoncelli, E. P. (2009). Is the homunculus "aware" of sensory adaptation? *Neural Computation*, *21*(12), 3271–3304.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, *90*(22), 10749–10753.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633. doi: 10.1371/journal.pcbi.1006633

Siegel, S., Allan, L. G., Hannah, S. D., & Crump, M. J. C. (2009). Applying signal detection theory to contingency assessment. *Comparative Cognition & Behavior Reviews*, *4*, 116–134.

Soto, F. A. (2019). Beyond the "conceptual nervous system": Can computational cognitive neuroscience transform learning theory? *Behavioural Processes*, *167*, 103908. doi: 10.1016/j.beproc.2019.103908

Soto, F. A., & Narasiwodeyar, S. (2021). Improving the validity of neuroimaging decoding tests of invariant and configural neural representation. *bioRxiv*, 2020.02.27.967505. doi: 10.1101/2020.02.27.967505

Soto, F. A., Stewart, R. A., Hosseini, S., Hays, J. S., & Beevers, C. G. (2021). A computational account of the mechanisms underlying face perception biases in depression. *Journal of Abnormal Psychology*, *130*(5), 443–454.

Soto, F. A., Vucovich, L. E., & Ashby, F. G. (2018). Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data. *PLoS Computational Biology*, *14*(10), e1006470.

Sprague, T. C., Adam, K. C. S., Foster, J. J., Rahmati, M., Sutterer, D. W., & Vo, V. A. (2018). Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. *eNeuro*, *5*(3), ENEURO.0098–18.2018. doi: 10.1523/ENEURO.0098-18.2018

Sprague, T. C., Boynton, G. M., & Serences, J. T. (2019). The importance of considering model choices when interpreting results in computational neuroimaging. *eNeuro*, *6*(6), e0196–19.2019. doi: 10.1523/ENEURO.0196-19.2019

Sprague, T. C., & Serences, J. T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature Neuroscience*, *16*(12), 1879–1887. doi: 10.1038/nn.3574

Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, *23*(8), 775–785.

Turner, B. M. (2015). Constraining cognitive abstractions through Bayesian modeling. In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 199–220). New York: Springer.

Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206. doi: 10.1016/j.neuroimage.2013.01.048

Turner, B. O., Mumford, J. A., Poldrack, R. A., & Ashby, F. G. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage*, *62*(3), 1429–1438. doi: 10.1016/j.neuroimage.2012.05.057

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.

Van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). New York: Cambridge University Press.

van Gerven, M. A. J. (2017). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, *76*, 172–183.

Vazquez, A. L., & Noll, D. C. (1998). Nonlinear aspects of the BOLD response in functional MRI. *NeuroImage*, *7*(2), 108–118.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. doi: 10.1016/j.neuroimage.2015.12.012

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience*, *29*(34), 10573–10581.

Wandell, B. A., & Winawer, J. (2015). Computational neuroimaging and population receptive fields. *Trends in Cognitive Sciences*, *19*(6), 349–357.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, *110*, 48–59.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, *10*(2), 403–430.

Zucchini, W. (2000, March). An introduction to model selection. *Journal of Mathematical Psychology*, *44*(1), 41–61.

Index