Julia Riebel · Hannah Lichtenberg

Formative Modelling in Psychology and Educational Science

Data-driven Index Formation According to the MARI Method

Springer

Formative Modelling in Psychology and Educational Science

Julia Riebel • Hannah Lichtenberg

Formative Modelling in Psychology and Educational Science

Data-driven Index Formation According to the MARI Method

Springer

Julia Riebel
Institute of Educational Science
Chair of School Pedagogy and Empirical Educational Research
RWTH Aachen University
Aachen, Germany

Hannah Lichtenberg
Institute of Educational Science
Chair of School Pedagogy and Empirical Educational Research
RWTH Aachen University
Aachen, Germany

ISBN 978-3-658-39403-5 ISBN 978-3-658-39404-2 (eBook)
https://doi.org/10.1007/978-3-658-39404-2

This Springer imprint is published by the registered company Springer Fachmedien Wiesbaden GmbH, part of Springer Nature.
The registered company address is: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Preface

What you can't describe well, you can't measure. (René Descartes)

Latent constructs, as they occur in test or questionnaire procedures in research and practice in educational science and psychology, "are not simply there and only have to be made observable, but have to be constructed first, building on what can be observed" (Steyer and Eid 2001, p. 4).

In this context, the view of the latent construct, which is in the background and causally causes the response of manifest indicators (e.g., questionnaire items), is only one of two possible perspectives. The item here reflects the underlying construct, which is why we also speak of the reflective measurement model here.

Formative measurement models, on the other hand, describe a different, constructivist view of latent constructs, which, as an alternative to the reflective approach, assumes that some constructs only emerge through the interaction of various and different indicators and are thus causally (i.e., temporally or logically) subordinate to them.

In the behavioral sciences, such formative models are largely unknown, although their application could be profitable here as well (Jarvis et al. 2003). The modeling of formative measurement models will be given space in this book, so that users in education and psychology get to know and apply the alternative to the reflective measurement model. This helps not only in the development of new measurement procedures but also in the modeling of already-known scales, which can be subjected to a critical second look with the knowledge of formative models. This can expand the knowledge of constructs and their measurement.

The first three chapters of this book present and discuss key principles about formative models. Chapter 1 describes the logic of the formative measurement

model in distinction to the reflective variant. Chapter 2 is dedicated to the question of how users can recognize whether formative models can be applied to a certain construct and shows possibilities for the practical implementation of such models. Based on critical aspects of formative models, Chap. 3 discusses their potential for application in educational science and psychology.

Chapter 4 describes a separate approach to the implementation of formative measurement models. The MARI method described here (Fluck 2020b) is used to implement data-driven index formation.

M	Mental experiments are objectified by expert judgments and ultimately determine whether the formative model is appropriate.
A	Analyses of the item designs, based on both qualitative and quantitative data, are used to adjust the scales.
R	Regression analyses additionally support the examination of scale quality by modeling critical validity.
I	Index formation is based on the information obtained in advance and is thus theoretically and empirically sound.

It is particularly important to supplement quantitative data with qualitative data from expert interviews. The integration of such data is central, as the quality of constructs with formative indicators cannot be assessed on the basis of individual key figures alone. In order to be able to calculate and assess formative models, a differentiated view of the construct and the process of test construction is necessary (Albers and Hildebrandt 2006).

This book is intended to provide readers with the necessary basic knowledge of formative models and to enable them to calculate such models using standard methods. In this way, we hope to contribute to a broader knowledge and use of this approach in educational science and psychology.

Aachen, Germany — Julia Riebel
Aachen, Germany — Hannah Lichtenberg
October 2020

Contents

The Logic of the Formative Measurement Model

1

Measuring something that is not directly observable presents us as researchers with a particular challenge. We acknowledge that intelligence, motivation, learning success, satisfaction, etc. are not accessible to direct sensory experience and look for ways to make the impossible (approximately) possible and still capture these constructs empirically.

To build this bridge, we need measurement models that formalize the relationship between latent (i.e., unobservable) constructs and their indicators, which are supposed to capture them approximately. Such indicators can be, for example, individual tasks or questions (items) in a questionnaire.

The literature distinguishes between two types of measurement models, the formative and the reflective measurement model, which are described and differentiated from each other in this chapter. In the research literature, there is a pronounced "dominance of reflective measurement models" (Fuchs 2011, p. 9), which should be familiar to the readers, which is why the special features of the formative measurement model are discussed here in particular.

1.1 Background: Latent Variables

"I only believe in what I see, what I can put on the table in front of me and touch," says a student. The religion teacher smiles superiorly and asks the student to please put a "pound" of his intelligence on the table.

The teacher, however, has not succeeded in proving God through this rhetorical maneuver. The intelligence of the student may not be observable or tangible at first

J. Riebel, H. Lichtenberg, *Formative Modelling in Psychology and Educational Science*, https://doi.org/10.1007/978-3-658-39404-2_1

glance. As a latent construct, it is not directly accessible. This does not mean, however, that it *cannot be made* observable and even quantifiable.

In order to make latent constructs such as intelligence measurable, it is necessary to operationalize them using concrete, manifest items in the form of questions or tasks within a test. For example, the above-mentioned student could "put on the table" a series of correctly answered intelligence items from a standardized test procedure.

However, intelligence is not directly visible through this either. One can only *infer* the underlying level of intelligence from the answers to the tasks. It is not for nothing that one speaks here of the bridging *problem*, for there is a gap to be bridged (Steyer and Eid 2001, p. 2) between what can neither be observed nor measured in the strict sense and the items that purport to do just that. Thus, the resourceful teacher might doubt whether the intelligence tasks the student has worked on measure his or her intelligence at all, and if so, how well. This brings up the two critical elements in test construction, the psychometric criteria of reliability and validity (Jäger and Petermann 1995).

Latent constructs are often measured in research not only for their own sake, but usually with the ultimate goal of investigating correlates with other constructs with which complex cause-effect relationships exist. Such causal relationships are mapped in linear structural equation models (hereafter abbreviated SEM after structural equation modeling), in which the relationships between multiple latent variables are "formally framed in such a way that their validity can be subjected to empirical testing" (Weiber and Mühlhaus 2014, p. 3). This is done by "drawing inferences from the empirically measured variances and covariances […] as to which dependency relationships exist between the underlying latent variables" (Fuchs 2011, p. 2). At the same time, the quality criteria of reliability and validity can also be tested in the context of test development using structural equation models.

The structural equation model represents a combination of confirmatory factor analysis (measurement model, relationships between the manifest variables) and regression or path analysis (structural model, relationships between the latent variables) (see Fig. 1.1).

In the structural model, the active structure between the latent constructs is mapped. The relationships that are defined here must either be derived from theory or be factually logical. If the regression coefficients between the latent constructs are in conformity with the theory, this fact serves to prove (convergent[1] or

[1] Convergent validity can be demonstrated by expectedly high correlations with related constructs, discriminant validity by expectedly low correlations with distinct constructs.

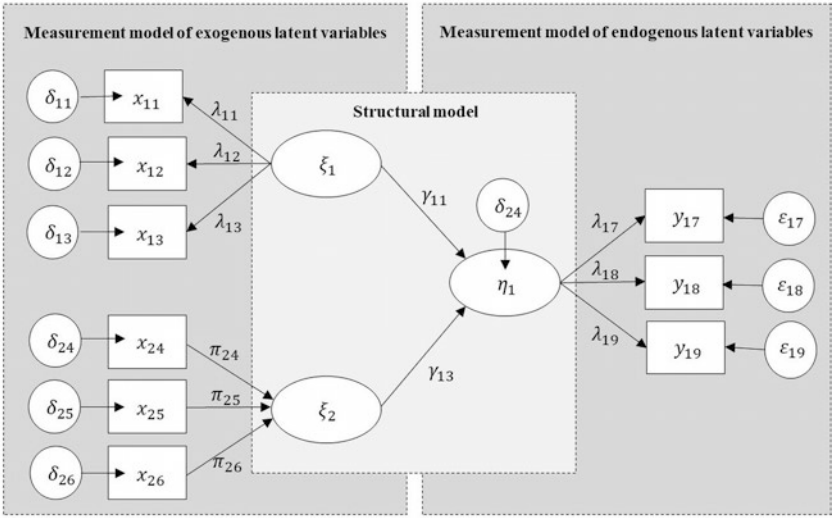


Fig. 1.1 Example of a linear structural equation model with one dependent (endogenous) and two independent (exogenous) variables. This structural equation model represents a combination of confirmatory factor analysis (measurement model) and regression analysis (structural model). (Own illustration based on Fuchs 2011)

discriminant) validity. The relationships are represented in systems of linear equations. SEM "are used to estimate the coefficients of effect between the variables under consideration and to estimate the measurement error" (Weiber and Mühlhaus 2014, p. 7). However, the correlative relationships found are only necessary, not sufficient conditions for causality (Fuchs 2011). The postulated causal relationships cannot be confirmed with SEM either, but only falsified.

The measurement model serves to bridge the gap between theory and empiricism and is tasked with answering the question, "How can one link an empirical theory and the theoretical concepts it contains to observables?" (Steyer and Eid 2001, p. 2). Therefore, in the measurement model, latent variables are defined. They must be assigned "suitable empirical indicators that describe the latent construct as accurately as possible" (Fuchs 2011, p. 5). Such indicators are the individual items in tests or questionnaires. In a confirmatory factor analysis (CFA), which either forms part of the SEM or can alternatively be calculated in isolation, it must be shown that each indicator is significantly related to the construct and shares a sufficiently large proportion of variance with the other indicators. If this is the case, evidence for reliability in the sense of internal consistency is provided. A

theory-compliant factor structure that confirms the anticipated numbers of factors and affiliations to indicators also serves as an indication that construct validity is given.

In the measurement model, moreover, the relationship between indicators and constructs is defined. According to Steyer and Eid (2001), this concerns firstly the magnitude of the relationship, which can be captured with a simple factor loading, but secondly also the question of exclusivity: Does each item load on only one factor (= does it measure only one construct) or do significant secondary loadings exist (= does the item measure more than one construct)? For example, mathematical text tasks measure language and reading skills in addition to mathematical skills (Riebel 2010). Third, however, there is also the question of the direction of the relationship between item and construct and thus the question of formative versus reflective measurement models to be addressed in the next section.

1.2 Formative and Reflective Measurement Models

In psychology as well as in educational science, research questions often deal with a latent variable/construct which is to be made measurable by manifest indicators. Two variants are available for this, formative and reflective modeling. However, in practice the "assumption of reflectivity of indicators is usually not questioned" (Albers and Hildebrandt 2006, p. 3) and accordingly reflective measurement models are almost exclusively used. However, the appropriateness of this frequent use of the reflective measurement model is much debated in other disciplines, such as economics (see Albers and Hildebrandt 2006; Bollen 1989; Diamantopoulos and Winklhofer 2001; Eberl 2004).

Reflective and formative measurement models differ fundamentally in their assumptions of causality. The reflective measurement model is an expression of a view in which the construct also exists independently of the observer and the observation process. The formative model, on the other hand, is well compatible with a constructivist perspective in which the construct as such only comes into being in the particular form by being measured (Borsboom et al. 2003). Using the right measurement model is essential because misspecification can lead to erroneous results (see Albers and Hildebrandt 2006; Bollen 1989; Diamantopoulos and Winklhofer 2001; Eberl 2004). The next section will first elaborate on the differences between reflective and formative measurement models. Consequences of flawed modeling will be addressed in Sect. 3.2.

While psychology and educational science traditionally work a lot with reflective measurement models, formative measurement models have been numerous in

the economic sciences for a long time. One reason for this may be the disparity of the phenomena under consideration and their causality assumptions. Reflective measurement models describe constructs in which various indicators reflect the latent construct. These are behavioral constructs that are not directly observable and are reflected equally in all their indicators (Albers and Hildebrandt 2006). A typical example from psychology is intelligence, which as a latent construct is not directly measurable, but can be made measurable using tasks in intelligence tests (Bollen 1989). Intelligence as an underlying construct has an influence on the extent to which various tasks are solved correctly. Conversely, however, intelligence is defined independently of the specific tasks.

Formative measurement models, on the other hand, describe constructs in which the variable to be explained is composed of various indicators, each of which describes different aspects of the construct. A typical example from economics is socioeconomic status (SES), which is composed of various indicators describing life circumstances, e.g. income, occupation and housing conditions (Hauser 1971).

In Fig. 1.2, these differences can be seen from the direction of the arrows. In a reflective measurement model, the arrowheads point from the latent construct (intelligence) to the indicators (tasks), i.e. intelligence is understood as causally prior to the tasks. In a formative measurement model, the arrows point from the indicators (life circumstances) to the latent construct (SES). In the literature, reflective indicators are sometimes referred to as "effect indicators," the indicators show the effect that the latent construct has on something, that is, changes in the construct are reflected in changes in all indicators (Christophersen and Grape 2009). Formative indicators consist of the causes of the latent construct and are therefore sometimes referred to as "cause indicators".

Even if some constructs suggest a certain type of modeling due to their definition, it should nevertheless be noted that a construct usually cannot be modeled formatively or reflectively per se, but the type of specification depends on the respective type of operationalisation. An illustrative example is described by Albers

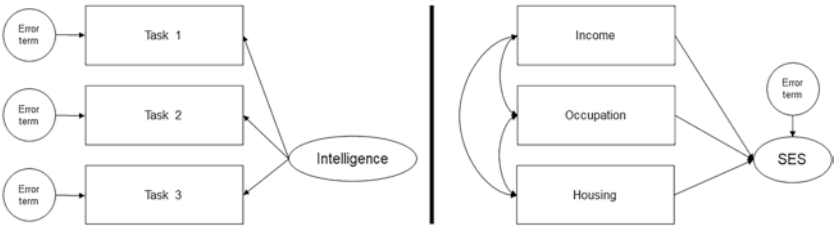


Fig. 1.2 Reflective (left) and formative (right) measurement model (Fluck 2020a)

and Hildebrandt (2006) using the example of satisfaction. Transferred to the pedagogical context (Funke et al. 2020), satisfaction can be a formative construct if the different causes of satisfaction – for example with a study program – are measured by different indicators ("I am satisfied with the study plan"; "I am satisfied with the professional competence of the lecturers", …). These indicators are causally prior to the construct to be measured and in combination lead to satisfied students. In order to adequately measure the construct, all aspects that contribute to satisfaction must be operationalized and surveyed. Satisfaction, on the other hand, is a reflective construct when the causally subordinate consequences of satisfaction as a psychological construct are inquired about, i.e. how satisfaction affects the level of experience and behavior ("I would recommend the study program to another person"; "I would choose this study program again").

The main differences between the two measurement models are summarized in Table 1.1 and explained in more detail below.

Table 1.1 Differences between reflective and formative measurement models (cf. Fluck 2020a)

Features	Reflective measurement model	Formative measurement model
Model equation	$X_i = \lambda_i \eta + \varepsilon_i$	$\eta = \sum Y_i X_i + \zeta$
Measurement error	Item-level measurement error	Measurement error at the construct level
Importance of the indicators	Samples from an item universe	Distinct aspects of the construct
Interchangeability of indicators	Items are interchangeable	Items are not interchangeable
Meaning of the removal of an indicator	Reduction in reliability	Change in the meaning of the construct
Correlation of the indicators	Theoretical perfect intercorrelations between items reduced only by measurement error	Indicators can correlate positively, negatively or to zero
Scale adjustment	Reliability analyses	Expert judgments, External criteria, Collinearity analysis
Model identification	Isolated CFA from 3 indicators identified	Isolated CFA fundamentally underidentified
Review of the measurement model	CFA	MIMIC model

Model Equation and Measurement Error

In the reflective measurement model, the indicators represent the dependent variables because they are predicted by the construct. The items (x_i) are described by a linear function of the latent construct (η), the respective load (λ_i) and a measurement error (ε_i),

$$X_i = \lambda_i \eta + \varepsilon_i$$

The measurement error here describes the proportion of variance in the item that cannot be explained by the latent construct. In the case of perfectly reliable items, the measurement error is $\varepsilon = 0$.

In the formative measurement model, the indicators (x_i) do not have a measurement error, since they are independent variables and are not explained by the model. The dependent variable of the formative model is the latent construct (η) to which an error (ζ) is assigned. The latent construct is modeled by a regression, as each indicator is weighted with its respective regression coefficient (γ_i) in the equation (Diamantopoulos and Riefler 2008).

$$\eta = \gamma_i X_i + \gamma_2 X_2 + \ldots + \gamma_n X_n + \zeta$$
$$\eta = \sum \gamma_i X_i + \zeta$$

The error term ζ is not understood as measurement error per se. The error term can be understood as all those aspects that are included in the latent construct but are not represented by the indicators of the model (Diamantopoulos 2006). If one assumes that a construct is composed of different facets (each measured by its own indicators), then it is no longer a question of the measurement error with which the individual indicators are provided in the sense of classical test theory. Rather, the quality of the measurement is measured by the extent to which the construct is more or less completely covered by the indicators. Another interpretation is that of Jarvis et al. (2003), for example, who interpret the measurement error as a collective measurement error across all indicators.

Importance of Interchangeability and Removal of an Indicator

While in a reflective model the different indicator items are interchangeable, this is not the case in the formative measurement model. Referring to the previous examples, equally difficult tasks can be exchanged without the measured latent construct, intelligence, changing in its expression. However, if one exchanges the indicator item occupation for another item such as education, the construct of

socioeconomic status now says something different. The same applies if an item is omitted. If an indicator item is omitted in the reflective measurement model, this only affects reliability. However, if an indicator item is removed in the formative measurement model, this can have a major impact on the validity of the model, as it now no longer depicts the same thing, but has a different meaning in terms of content.

Correlation of the Indicators

The indicator items of a reflective measurement model are conceptually samples from an imaginary item universe; they measure the same thing, only differently. The correlation between the different items is correspondingly high, since they are all influenced by the same latent construct. To stick with the intelligence example: The probability that a person who correctly solved tasks 1 and 2 will also correctly solve task 3 is high, since intelligence as a stable construct influences the answering of all questions. If the measurement error of all indicator items were 0 ($\varepsilon = 0$), this would correspond to a perfect correlation of all indicators (Eberl 2004). Thus, one criterion for the goodness of a reflective measurement model is a high correlation of the items (Bollen and Lennox 1991). In the formative model, there is no factual logical justification for such correlation. Unlike the reflective model, the expectation of high correlation of indicators does not apply here. The correlation of indicators in the formative measurement model can be high, low, zero, or even negative. A good example of this comes from stress research. Life circumstances such as high job demands, recent death of a loved one, etc., can be indicators of a person's overall stress. While it is possible that a person who reports suffering from the high demands of their job also reports that a relative has died, in many cases this combination is unlikely to apply (see Edwards and Bagozzi 2000).

Scale Adjustment and Construct Validation

Methods of scale adjustment in classical test theory are based on the fact that the items are understood as samples from an item universe and are highly correlated in the case of reliable measurement. Reliability analyses and confirmatory factor analyses (CFA) are therefore only suitable for construct validation in the case of reflective models.

On the other hand, other methods must be used to test the formative model, as the correlations between indicators are not relevant here. In this case, expert judgements, external criteria and collinearity analyses are used. Since the formative measurement model corresponds to a regression function, multicollinearity is to be avoided since otherwise the regression coefficients may no longer be unambiguously determinable and the validity testing of the indicators becomes problematic

(Eberl 2004). The Multiple Indicators and Multiple Causes (MIMIC) model (Jöreskog and Goldberger 1975) is often used to test the measurement model. The various methods for validating the formative measurement model are discussed in more detail in Chap. 2, and details on scale adjustment are described in Sect. 4.2.

Model Identification

In the context of structural equation models, we speak of identification when sufficient empirical information is available to unambiguously estimate all parameters of the model (Weiber and Mühlhaus 2014). The model is uniquely estimable and solvable only if the necessary condition for identifiability, called the t-rule, is satisfied. The t-rule states that the number of variances and covariances of the empirically collected data must be at least equal to the number of model parameters to be estimated (t):

$$t \leq \frac{1}{2}(p+q)(p+q+1)$$

p stands for the number of dependent indicators and q for the number of independent indicators (Temme 2006). This means that if estimated values can be calculated for all unknown model parameters (t), the overall model is considered identifiable[2] (Weiber and Mühlhaus 2014).

Viewed in isolation, a reflective measurement model is identified from three indicators, whereas a formative model is fundamentally underidentified. In Fig. 1.2, for example, the reflective model is just identified, it has six indicators to be estimated (three loadings and three measurement errors) and six available pieces of information (three variances and three covariances) (Eid et al. 2015). In contrast, in the formative measurement model shown in Fig. 1.2, seven parameters would have to be estimated from the available six pieces of information (three intercorrelations of the indicators, three regression weights, and one latent error term).

As a consequence, methods such as CFA cannot be applied to formative models. There are several ways in which the identification problem can be solved in the context of structural equation modeling. These are described in Chap. 2.

[2] This is at least theoretically the case; in the case of multicollinearity of indicators, a model may in fact be underidentified despite the t-rule.

The Application of the Formative Measurement Model

2

Prior to considering *how* a construct is formatively modeled, it is first necessary to decide *whether* formative modeling is appropriate or whether a reflective construct better describes theory and/or empiricism. This question is addressed in the first Sect. 2.1 of this chapter. Only when this question has been answered satisfactorily can the implementation of the measurement model begin. Section 2.2 presents different approaches to formative modeling, including variance- and covariance-based implementations of different modeling variants such as two-construct models, nomological networks, and MIMIC models. Finally, the data-guided index formation presented in this book is discussed in Chap. 4.

2.1 Formative or Reflective Modeling?

Is a given construct to be modeled formatively or reflectively? There are two positions on the question of how to decide this. In principle, there is agreement that this decision must be made a priori, i.e. it must be decided at the beginning of the test or questionnaire development whether a formative or reflective model is the appropriate one. This decision must be made on the basis of theoretical and logical considerations (Sect. 2.2.1). Some authors also advocate empirical testing of appropriateness. However, this should be done in *addition* to, and not as an alternative to, theoretical considerations, since the question of the correct specification cannot be answered from empirical evidence alone (Sect. 2.2.2).

However, as already illustrated by the example of satisfaction, the question of specification cannot be answered for a construct per se, but always for a concrete type of operationalization (Albers and Hildebrandt 2006).

J. Riebel, H. Lichtenberg, *Formative Modelling in Psychology and Educational Science*, https://doi.org/10.1007/978-3-658-39404-2_2

2.1.1 Decision Based on Theory and Factual Logic

In principle, the determination of the type of specification should be done a priori (Bollen 1989). This determination is made depending on the theory and conceptualization of a construct (Fornell and Cha 1994). It is possible that the existing literature provides insight into the relationship between indicator and construct. For example, this is the case in parts of the research literature on school violence or workplace bullying (Festl 2015; Tepper and Henle 2011). Here, there are references in several publications to the fact that the constructs in question should be modeled formatively (cf. Fluck 2020a), even if in practice there are only isolated studies in which this modeling is then also implemented.

Constructs such as socioeconomic status (Hauser 1971) are already considered formative in previous work – and accordingly already modeled as an index. Such preliminary work provides an argumentative basis for formative modeling.

For other constructs, the evidence in the research literature may be more subtle. In health psychology, the construct of social support is described as something that is additively composed of various social resources. Although formative models are not explicitly mentioned here as a method of choice, the relevant thought process lies behind them (Taylor 2011).

In the absence of any such clues about a construct from the literature, Bollen (1989) suggests approaching the direction of causality between item and construct with mental experiments. These mental experiments are used to elicit the correct specification based on guiding questions. Usually, the following three questions are used (Eberl 2004; for details on the guiding questions see Sect. 4.1):

1. **Do the items all measure the same thing, i.e. are they interchangeable?** If the construct changes in meaning as soon as an item is omitted, this speaks in favor of formative modeling, since the items then obviously do not all measure the same thing.
2. **What is the direction of causality?** Is the construct in the background and causes the way subjects react to individual test and questionnaire items (reflective) or does the construct only come together through the combination of indicators (formative)?
3. **What consequences result from changes?** If we assume a change in the construct, a reflective measurement model should also result in changes in all indicators.

The following example illustrates the application of the guiding questions to a concrete phenomenon.

Guiding Questions for the Specification of Cyberbullying

The construct of cyberbullying is often recorded in questionnaires in such a way that various forms of violence on the Internet or via mobile phone are named and the affected persons are asked to tick which of these forms they have experienced. Festl (2015) and Fluck (2020a) argue on the basis of these guiding questions that cyberbullying should be regarded as formative in the following operationalization:

How often in the past year (never – only once or twice – about once a month – about once a week – several times a week) have you been attacked, insulted, threatened or humiliated by your classmates through the following media:

- *Text messages*
- *Emails*
- *Calls*
- *Chats*
- *Instant messenger*
- *Websites*

Question 1: Interchangeability of items

The items describe different ways of communication. Unlike other operationalization approaches that describe different types of attacks (insulting versus spreading rumors versus excluding from common activities, etc.), the items do not measure distinct behaviors. Yet they are not interchangeable in the sense that they measure the same thing. If one were to change the questionnaire and delete some of the communication channels, the construct being measured would likely change as well. However, the first question cannot be answered clearly on the basis of the existing preliminary information.

Question 2: Direction of causality

Which came first, the item or the construct? The assignment of the roles of hen and egg are often circular and the question of causality difficult to answer. It is often possible to argue both ways, even in the case of the construct of intelligence, which has already been described several times as typically reflective, the well-known quote by Boring (1923) that intelligence is what the intelligence test measures seems to challenge this view. The question of causality can therefore be better answered by breaking it down to the direction of action between indicator and construct shown by the arrows in the measurement model:

(*continued*)

(continued)

In order to make the "diagnosis" (Fluck 2020a) that someone is a victim of cyberbullying, this person must have been attacked several times or over a longer period of time via new media without being able to defend themselves against it. These attacks can, but do not necessarily have to, occur using *different* forms and through *different* communication channels. The construct is therefore defined additively by the indicators. A high score on one indicator does not necessarily go hand in hand with a high score on another indicator. Thus, a victim of cyberbullying is not automatically likely to have a high score on all items.

Question 3: Consequences of changes

The third question clearly suggests formative modeling. A change in the construct should be reflected in all indicators if the reflective model is valid. If, for example, a person's ability to concentrate increases following training, it is to be expected that the items of a concentration test will all be solved with increased probability. This consequence is not necessary in formative modeling. For example, if we assume in the present example that a person suddenly experiences more victimization than before, it is quite possible that this occurs only on the basis of one medium and not necessarily via several channels as depicted in the various items.

The more questions underlie mental experiments, the clearer the picture researchers get of their constructs. Since the answers to the guiding questions can vary, as the example shows, the decision between reflective and formative modeling is not always clear-cut. The choice and weighting of the decision questions thus make the commitment to a specification subjective to some extent. This point is highlighted as particularly problematic by critics of the formative method (see Sect. 3.1).

2.1.2 Decision Based on Empirical Data

Due to the subjectivity of mental experiments described above and the criticism that follows, some authors take a different position and argue that the decision should also be based on empirical data. Although a hypothesis is to be made at the beginning on the basis of mental experiments as to which is the correct type of

specification, this hypothesis is subsequently to be confirmed on the basis of empirical investigations. However, a decision based *purely* on data is not envisaged in this approach either.

TETRAD Test

Bollen and Ting (1993, 2000) developed the confirmatory TETRAD test, which can be used to reject the reflective measurement model. Rejection of the reflective model, in turn, can be taken as an indication of the validity of the formative model. Based on the basic assumptions underlying a reflective measurement model, the test tests whether the conditions for this model are met by examining the covariances of the individual items for their interrelationships. Ultimately, therefore, the TETRAD test is a procedure that tests whether the assumption of intercorrelation of indicators underlying the reflective model is satisfied to a sufficient degree. The box below presents the principle of the test procedure (taken from Fluck 2020a).

Overview

The logic of the TETRAD test builds on the fact that the covariance of two items loading on a common factor can be expressed as:

$$\operatorname{cov}\left(x_1,x_2\right)=\lambda_1\lambda_2\phi,$$

Where λ_i denotes the loading of the i-th item on the latent construct ξ and ϕ denotes the variance of ξ.

This relationship is not obvious at first glance, but it can be understood on the basis of some assumptions from covariance algebra (these can be found, for example, in Bollen (1989)): For the covariance of two random variables X_1 and X_2 the following rules apply in connection with a constant c:

$$1.\ \operatorname{cov}\left(X_1,X_1\right)=\operatorname{var}\left(X_1\right)$$

$$2.\ \operatorname{cov}\left(c\,X_1,X_2\right)=c\,\operatorname{cov}\left(X_1,X_2\right)$$

$$3.\ \operatorname{cov}\left(X_1,X_2+c\right)=\operatorname{cov}\left(X_1,X_2\right)$$

The covariance of a variable with itself corresponds to its variance (1). If a variable X_1 is multiplied by a constant and then covariated with a second variable X_2, the result corresponds to the product of the constant and the covariance of both variables (2). If a constant is added to one of two variables, this has no effect on the covariance of the two variables (3).

The basic equation of the LISREL[1] -structural equation model states that each manifest variable x_i is a linear function of the latent variable ξ, weighted by its respective loading coefficient λ_i and augmented by its measurement error δ_i. Therefore, for two variables x_1 and x_2, which load on the same factor, (Bollen 1989) holds:

$$x_1 = \lambda_1 \xi + \delta_1 \quad \text{and} \quad x_2 = \lambda_2 \xi + \delta_2$$

For their covariance, that is:

$$\begin{aligned} \operatorname{cov}(X_1, X_2) &= \operatorname{cov}(\lambda_1 \xi + \delta_1, \lambda_2 \xi + \delta_2) \\ &\quad | \text{Prerequisite}(3) \\ &= \operatorname{cov}(\lambda_1 \xi, \lambda_2 \xi) \\ &\quad | \text{Prerequisite}(2) \\ &= \lambda_1 \lambda_2 \operatorname{cov}(\xi, \xi) \\ &\quad | \text{Prerequisite}(1) \\ &= \lambda_1 \lambda_2 \phi \end{aligned}$$

Bollen and Ting (1993) define for four indicators x_1 -x_4 the so-called tetrad τ_{1234} as the difference between the product of two covariances and the product of two other covariances:

$$\tau_{1234} = \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24}$$

For four random variables, there are two additional ways to combine the covariance pairs, namely:

$$\tau_{1342} = \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{32} \quad \text{and}$$
$$\tau_{1423} = \sigma_{14}\sigma_{23} - \sigma_{12}\sigma_{43}$$

One of the basic assumptions of the reflective measurement model in the LISREL approach is the relationship between indicator covariances and loadings and variance of the latent construct derived above:

$$\sigma_{12} = \lambda_1 \lambda_2 \phi.$$

[1] LISREL does not stand for the concrete program here, but for all covariance-based estimation methods in the framework of “linear structural relations”.

Substituting this into the defining equation of the tetrads, we get:

$$\begin{aligned}\tau_{1234} &= \lambda_1\lambda_2\phi \cdot \lambda_3\lambda_4\phi - \lambda_1\lambda_3\phi \cdot \lambda_2\lambda_4\phi \\ &= \lambda_1\lambda_2\lambda_3\lambda_4\phi^2 - \lambda_1\lambda_2\lambda_3\lambda_4\phi^2 \\ &= 0\end{aligned}$$

Thus, insofar as the reflective measurement model holds, all possible combinations of tetrads must be equal to zero; Bollen and Ting (1993, 2000) therefore speak of "*vanishing tetrads*". If the tetrads do not "vanish", in other words if some or all of the tetrads are significantly greater than zero, then the reflective measurement model must be rejected. The TETRAD test is thus a test for the validity of the reflective measurement model.

With more than four indicator items, for n indicators, n!/(n − 4)!4! sets of tetrads can be formed, all of which must be tested for their difference from zero. For fewer than four indicators, the TETRAD test cannot be performed, or only by "provisionally" adding a fourth item from another scale, which, however, again affects interpretability (Gudergan et al. 2008).

Bollen and Ting (1993) developed a statistical test that, for a given empirical covariance matrix and the associated covariance matrix implied by the reflective model, simultaneously tests its significant deviation from zero for all possible tetrads. The test variable T^2 is approximately chi-squared distributed, where the number of degrees of freedom corresponds to the number of vanishing tetrads to be tested.

In practice, the TETRAD test can be implemented in several ways:

1. **Implementation in SAS** A SAS macro developed by Hipp et al. (2005) is available for the application of the TETRAD test. The procedure is described in detail in Hipp (2008), and an application example can be found in Bollen et al. (2009). To perform the TETRAD test in SAS, the empirical variance-covariance matrix and the ("true") variance-covariance matrix implied by the reflective measurement model are compared. The model-implied matrix must be computed and read out in advance using Mplus or LISREL based on a CFA. An application example can be found in Fluck (2020a).
2. **Implementation in smartPLS** The program smartPLS (Ringle et al. 2015), which was developed for the estimation of PLS models, allows a simple check of the tetrads by means of a few clicks. For each non-redundant tetrad, a test statistic T and an associated p-value as well as a confidence interval are calculated, among others. If one of the tetrads differs significantly from zero, the re-

[2] Detailed information on the calculation of the test variable T can be found in Bollen and Ting (1993, 2000).

flective measurement model can be discarded. Details on the logic and use of the TETRAD test in smartPLS can be found in Gudergan et al. (2008).

3. **Evasion to CFA** Finally, the TETRAD test examines whether the basic assumptions of the reflective measurement model are valid. In a comparison of the TETRAD test and confirmatory factor analyses, Binder and Eberl (2005) were able to show "that neither procedure provides information about the data structure that would not already be available from the other (p. 15)." Only with smaller samples, according to the authors of the study, does the TETRAD test tend to reject the reflective model more than the CFA.

Algorithm by Eberl (2004)

The procedure developed by Eberl (2004) for empirically testing the adequacy of the formative model goes far beyond the use of a TETRAD test. As Fig. 2.1 shows, the process begins with a specification hypothesis that is to be formed on the basis of the mental experiments described earlier. However, the resulting determination on a modeling method is considered provisional here and is to be tested in the further course using several empirical procedures.

First, a scale adjustment is carried out, in the formative variant by means of expert judgements or external validation (see Chap. 4). In addition to the TETRAD test, the construct should then be modeled both reflectively and formatively. In the reflective procedure this can be done with factor analyses, in the formative procedure with MIMIC models, or with the help of other constructs within the framework of structural equation models. For formative modeling, Eberl (2004) recommends implementation using the variant-based approach (PLS modeling) described in more detail in Sect. 2.2.3. Reflective models can be calculated using covariance-based approaches (LISREL[3] modeling).

Regardless of the concrete implementation of both modeling variants, the goal in each step is to confirm a model with the assumed specification hypothesis or to reject a model with the respective opposite specification hypothesis. Based on these three results – TETRAD test, formative implementation and reflective implementation – information about the correct specification should be obtained. Ideally, the procedures will not lead to conflicting results. As indicated by the arrows in the margin of the figure, the process should be stopped as soon as a result contradicts the specification hypothesis and it cannot be clearly decided on the basis of empirics whether the construct is formative or reflective. The path then goes back to the conceptual level, and the construct may need to be reconsidered and re-operationalized.

[3] LISREL does not stand for the concrete program here, but for all covariance-based estimation methods in the framework of "linear structural relations".

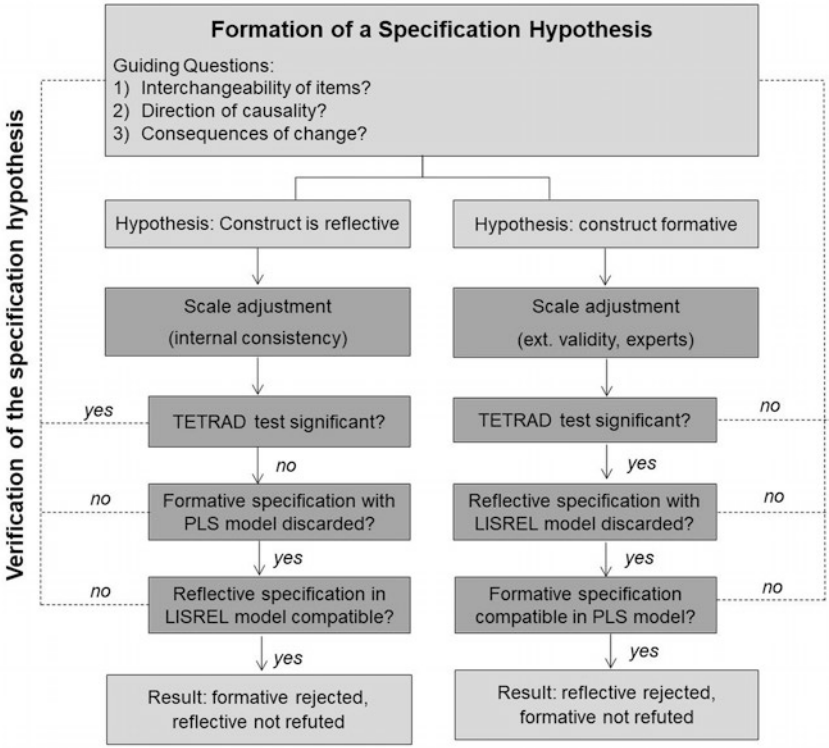


Fig. 2.1 Algorithm by Eberl (2004)

Evaluation of the Empirical Approaches

How can empirical approaches to specification be evaluated?

First of all, it must be emphasized at this point that an empirical test is always only useful as a supplement to factual considerations and cannot be carried out in the sense of an explorative analysis. If all assumptions regarding the question of how a construct is to be modeled are really missing, an expert survey is the better way to arrive at initial hypotheses.

Furthermore, it should be noted that discarding the reflective model (whether via TETRAD test, CFA or SEM) ultimately only answers the question of the intercorrelation of the indicators. If the reflective measurement model holds, then much (ideally: all) of the variance in the indicators must be explainable by the common underlying construct. Low intercorrelation among the indicators is inconsistent

with this assumption and leads to the rejection of the model. However, while inspection of the correlation matrix only provides clues, using TETRAD test and the model fit indices of the CFA, it is possible to clearly accept or reject the basic assumptions of the reflective model. Thus, the value of the empirical procedures lies primarily in obtaining a statistically validated statement about whether the indicators are correlated or not. Since the results of the TETRAD test are concordant with those of a single-factor CFA (Binder and Eberl 2005), it is sufficient to restrict oneself to one of the two procedures in order to reject the reflective modeling approach.

Nevertheless, the correlation of the indicators only allows us to say that the reflective model does *not* apply. In the formative model, the items do not *have to* correlate, but they *may* correlate. This creates the following problem: If one has a formative construct in which the indicators correlate, the TETRAD test and CFA may fail in favor of the reflective model. At the same time, a second type of fallacy is possible. The CFA may also reject a reflective model because a multidimensional construct was modeled unidimensionally or because other characteristics of the model were not appropriately co-modeled, such as the additional correlation of measurement errors for items that are similar to each other. However, this does not mean that the formative model is valid, only that the reflective model was misspecified.

Both wrong decisions can be countered by supplementing the empirical rejection of the reflective model with an attempt to confirm the formative model, as provided for in Eberl's algorithm. The fact that contradictory results can occur here only illustrates once more the importance of factual logic and theory-based considerations made in *advance*.

Since empirical procedures are often described in the literature in the context of critiques of the formative model, the suspicion sometimes arises that a number-based justification for the formative model is to be provided. However, it is difficult to see why an inductive procedure (which is particularly prone to error) should be fundamentally superior to a deductive procedure. Measures to increase the validity of a priori decisions on specification are presented in detail in the fourth chapter.

In summary, quantitative-empirical methods may not determine the decision-making process about the specification of a construct, but they can be valuable complementary inputs to expert judgments and theory-based reasoning.

2.2 Implementation of the Formative Model in Practice

Considered in isolation, formative measurement models cannot be estimated; confirmatory factor analyses, as are common with reflective measurement models, are not possible here. For modeling purposes, a formative model can either be implemented as an index (cf. Diamantopoulos and Winklhofer 2001; Christophersen and Grape 2009) or calculated using variance- or covariance-based procedures within the framework of structural equation models. The variance- and covariance-based approaches are not substitutive but complementary. While the covariance-based approach is theory-testing, the variance-based approach is data- and forecast-oriented (Weiber and Mühlhaus 2014).

Three models that can be used to build structural equation models are the MIMIC model, the nomological network and the two-construct model. These three models are described in more detail in the subsection on covariance-based approaches (Sect. 2.2.2). Section 2.2.3 then goes into more detail on the variant-based PLS method and what distinguishes this method from the covariance-based LISREL[4] approach.

Two examples of index formation are described in Sect. 2.2.1; another data-driven approach to index formation, the MARI method, is described in Chap. 4.

2.2.1 Index Formation

A common way of dealing with formative models is indexing. In this method, the formative indicators are combined into an index. Subsequently, the index is used for further calculations or modeling instead of the individual indicators. This is similar to combining reflective items into a scale (Bollen and Lennox 1991). There are several points to consider when constructing formative multi-item measurement models. Diamantopoulos and Winklhofer (2001) developed guidelines for constructing formative indices, which were taken up and slightly modified by Christophersen and Grape (2009). Both approaches will be briefly explained here (see Table 2.1).

Step one:

The first step for both authors is to determine the construct and define the latent variable. Since in a formative measurement model the indicators influence the la-

[4] LISREL does not stand for the concrete program here, but for all covariance-based estimation methods in the framework of "linear structural relations".

Table 2.1 Guidelines for the construction of formative indices

	Diamantopoulos and Winklhofer (2001)	Christophersen and Grape (2009)
Step 1	Content specification	Definition of the construct
Step 2	Indicator specification	Determination of indicators
Step 3	Indicator collinearity	Treatment of multicollinearity
Step 4	External validity	Estimation of the measurement model
Step 5	*N/A*	Index calculation

Source: Own representation based on Diamantopoulos and Winklhofer (2001) and Christophersen and Grape (2009)

tent variable, it is essential to pay close attention to which facets of the construct need to be considered (Christophersen and Grape 2009; Diamantopoulos and Winklhofer 2001). That is, the definition of the construct must be simultaneously as broad as necessary and "as precise as possible" (Christophersen and Grape 2009, p. 109). Literature research and qualitative methods such as expert interviews are suitable for this purpose (see also Sect. 4.1). For Christophersen and Grape (2009), this first step also involves deciding whether the model should be viewed formatively or reflectively.

Step two:

The construct definition determines the determination of the indicators in the second step of index building. The selected indicators must reflect all facets of the construct selected in the previous step (Christophersen and Grape 2009; Diamantopoulos and Winklhofer 2001). A review of the selected indicators is necessary and can be carried out, for example, through preliminary empirical studies (see also Sect. 4.2).

Step Three:

Since the equation of a formative measurement model is a multiple regression, multicollinearity is a problem (see box). Step 3 is therefore dedicated to checking the collinearity of the indicators for both authors (Christophersen and Grape 2009; Diamantopoulos and Winklhofer 2001).

In the case of high multicollinearity, an index can be formed. In this case, the various items are combined and treated as a single indicator.

An alternative to circumvent the problem of multicollinearity is to compute the measurement model as a PLS regression instead of an OLS multiple regression (Christophersen and Grape 2009, p. 112). More on PLS modeling follows in the related section in this chapter.

To avoid redundancy of individual indicators, Diamantopoulos and Winklhofer (2001, p. 272) suggest an exclusion of indicators that represent an almost perfect linear combination of other indicators. Christophersen and Grape (2009, p. 112), on the other hand, advise against this prior approach, since an elimination of indicators can also mean a loss of information.

Causes and Identification of Multicollinearity

The identification of multicollinearity and, if necessary, its treatment is an essential step in regression-based analyses. Multicollinearity occurs when different exogenous variables are correlated with each other, i.e. there is a linear dependence between them. This dependence contradicts the assumption of regression models that the independent indicators are also independent of each other and can lead to over- or underestimation of the regression coefficients. Even the signs of the coefficients can turn out differently than expected, as high variances and standard errors can also result from high multicollinearity (Schneider 2009).

For a regression equation with three exogenous variables

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

multicollinearity exists if at least one of the coefficients k_j is not equal to zero. In this case, the exogenous variable x_1 could be represented in a linear function of the other independent variables (Schneider 2009):

$$x_1 = k_1 x_2 + k_2 x_3 + e$$

Figure 2.2 shows three different examples of multicollinearity, the intersections of the circles represent the respective correlations. In practice, the complete absence of a correlation ($k = 0$, no multicollinearity) and also its perfection ($k = 1$, perfect multicollinearity) between x_1 and x_2 is rarely the case. Usually, partial multicollinearity is present, the greater the intersection between x_1 and x_2, the greater the multicollinearity (Schneider 2009). The causes of multicollinearity are manifold and often hidden, therefore a thorough examination is necessary for identification (Schneider 2009).

Multicollinearity does not necessarily affect the estimation quality of a construct. A look at "the representation of the standard error illustrates the problem that high estimation precision and regression coefficients with high standard errors are not mutually exclusive" (Steffen 1994, p. 8), but it also shows that low standard errors are also possible despite low estimation precision:

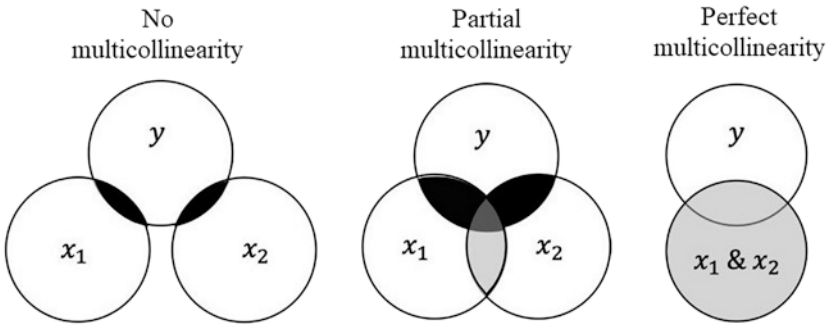


Fig. 2.2 Examples of multicollinearity. (Source: Own representation based on Schneider (2009, p. 222))

$$s_{b_1} = \left[\frac{s_e^2}{n s_{x_1}^2 \left(1 - B_{x_1 x_2}\right)} \right]^{1/2}$$

The standard error of the regression coefficient b_1 (s_{b_1}) is thus calculated from the estimation error of the regression (s_e^2), the number of observed values (n), the variance of the explanatory variables (s_x^2) and the multicollinearity between the corresponding explanatory variables ($B_{x_1 x_2}$). Thus, a high standard error can result from a high estimation error of the regression, a low number n, a low variance s_x^2, and high multicollinearity (Steffen 1994).

Other procedures are therefore necessary to identify multicollinearity. For regressions with only two predictors, a *covariance matrix* is useful. However, this is not sufficient for multivariate models, as low bivariate correlation coefficients do not necessarily mean that there is no multicollinearity due to the combined effect of several variables. Depending on how the indicators are scaled, the correlation coefficient is calculated differently. If the indicators are ordinally scaled, multicollinearity is tested using Spearman's rank correlation; if the indicators are metric, the Bravais-Pearson correlation coefficient is used (Moosmüller 2004, p. 22). A common threshold value is based on Cohen (1992), which states that values as low as .3 can be an indication of multicollinearity (Schneider 2009). Two other common methods are the calculation of the *tolerance level (TOL)* and the *variance inflation factor (VIF).* In these methods, the multiple correlation coefficient R_i^2 is first calculated by *auxiliary regressions of* all exogenous variables on each other. The tolerance value is calculated by.

$$TOL_i = 1 - R_i^2.$$

The smaller the tolerance value, the more likely it is that multicollinearity is present. The variance inflation factor (VIF) is the inverse of the tolerance value.

$$VIF_i = \frac{1}{TOL_i} = \frac{1}{1 - R_i^2}$$

Accordingly, high VIF values are indicative of multicollinearity (Schneider 2009).

There are different statements from various authors on the guideline values. A VIF > 10 can be considered particularly critical; items with VIF < 2 can be used without hesitation (Diamantopoulos and Riefler 2008; Schneider 2009; Weiber and Mühlhaus 2014).

In addition, various more complex numerical methods exist for the identification of multicollinearity (for an overview and critical classification see Schneider 2009).

Step four:

Since formative measurement models are never identified, it is also not possible to estimate the model parameters (see Sect. 1.2) and the model must be integrated into a larger model. In step 4 "construct validation", a two-construct model or a nomological network with formative and reflective indicators or a "multiple indicators and multiple causes model" (MIMIC model) are suitable (cf. Christophersen and Grape 2009; Diamantopoulos and Winklhofer 2001; Weiber and Mühlhaus 2014). All three models, which can also be calculated in the PLS variant, are described below.

Step Five:

In the fifth step, which is only explicitly mentioned by Christophersen and Grape (2009), the index is now calculated. The calculations of the mean value in index formation (see box) represent a disadvantage of this method, as condensation also means a loss of information (Albers and Hildebrandt 2006). In this way, the indicators are no longer considered individually, but collectively, and the question arises as to how exactly the index is to be interpreted. If the index formed is then used for further calculations and inserted into a regression function, a further question is what exactly is the meaning to be attributed to the associated regression coefficient (Diamantopoulos et al. 2008).

Advantages of indexing are the treatment of multicollinearity and the resulting possibility to work with regression analyses instead of having to incorporate the formative measurement model into more complex models.

Index Formation

There are various methods of combining indicators into an index. These include additive, multiplicative and weighted additive indexing. In the following and in Sect. 4.4, we will discuss weighted additive indexing in detail (for an insight into other methods, see e.g. Schnell et al. 2018).

Weighted additive index formation

For a weighted index (I), the indicators (x_i) are first multiplied by their weights (g_i) and then added up:

$$I = g_1x_1 + g_2x_2 + \ldots + g_nx_n$$

An important consideration here is the weighting of the individual indicators. This weighting can be done either individually per indicator or equally for all indicators. In the case of individual weighting, the correlations between the indicators and the dependent latent variable can be used as weights. Alternatively, a qualitative determination of the weights is also possible, which is based, for example, on existing literature or expert judgements. If the weighting is to be the same for all indicators, this can be done by calculating the mean.

Once the decision has been made that the weighting of the indicators should be quantitative, the way in which the index is calculated depends on whether the indicators compensate each other or not.

One speaks of a compensatory effect of the indicators if they balance each other out; if this is not the case, they behave non-compensatory. If, for example, the index is intended to represent academic success, the indicators "diligence" and "aptitude" can balance each other out (compensate) to a certain extent. In this context, however, the indicator "attendance at final examination" would have a non-compensatory effect – if I am not present at a final examination, I cannot compensate for this by being particularly gifted or diligent.

If the indicators are compensatory, either the respective regression weights can be used as weights (individual weighting) or an arithmetic mean can be formed from the normalized indicators (equal weighting for all indicators). If the indicators behave non-compensatory, only the geometric mean of the normalized indicators can be calculated; it is not possible to use the regression weights here (Albers and Hildebrandt 2006).

The reason for the two different calculations of the weights can be found in the formulas of the two means. While in the case of the geometric mean the index tends towards zero as soon as an individual indicator tends towards zero, since $\bar{x}_{geom} = \sqrt{\prod_{i=1}^{n} x_i}$, this is not the case with the arithmetic mean. $\bar{x} = \frac{1}{n}^{*} \sum_{i=1}^{n} x_i$

Referring to the previous example, it becomes clear here why the geometric mean must be chosen in the case of a non-compensatory effect:

$$g = \bar{x}_{geom} = \sqrt{x_{Final\ exam} \times x_{Aptitude} \times x_{Diligence}}$$

$$\text{if} : x_{Final\ exam} = 0$$

$$\text{then} : g = \bar{x}_{\text{geom}} = 0$$

It follows:

$$\text{I} = 0 \times x_{Final\ exam} + 0 \times x_{Aptitude} + 0 \times x_{Diligence}$$

$$\text{Accordingly} : \text{I} = 0$$

Literally speaking, if I do not take the final exam, my academic success is equal to zero, regardless of my diligence and aptitude. Using the regression weights is not possible for non-compensatory indicators, since the index must go to zero as soon as one of the indicators goes to zero.

However, if the index were calculated only from the compensatory indicators "aptitude" and "diligence", one could calculate academic success on the basis of the arithmetic mean of the two indicators.

$$I = gx_{Aptitude} + gx_{Diligence}$$

$$\text{Whereby:} \quad g = \bar{x} = \frac{1}{2}^{*} \left(x_{Aptitude} + x_{Diligence} \right)$$

My academic success in this case consists of the equally weighted indicators, my aptitude and my diligence.

2.2.2 SEM in the Covariance-Based Approach

Based on confirmatory factor analysis, the covariance-based approach simultaneously estimates all parameters of a structural equation model. The basis for the estimation is the empirical correlation matrix (or variance-covariance matrix), which is to be reproduced as accurately as possible with the help of a factor analysis (Weiber and Mühlhaus 2014). For the factor analysis, the latent variables are interpreted as factors, which are assigned to the measurement variables (indicators), according to the hypotheses. The correlation between the factors and measurement variables (factor loadings) are estimated in such a way that the model-theoretical correlation matrix represents as accurate a reproduction as possible of the empirical correlation matrix (Fuchs 2011; Weiber and Mühlhaus 2014).

The correlations between the *x* and *y* indicators shown in Fig. 2.3 form the basis for estimating the structural equation model (Backhaus et al. 2016).

The covariance-based approach offers several ways to estimate formative models: the two-construct model, a nomological network, and the MIMIC model. All three models have in common that the formative indicators are related to at least two reflective variables (see Fig. 2.4).

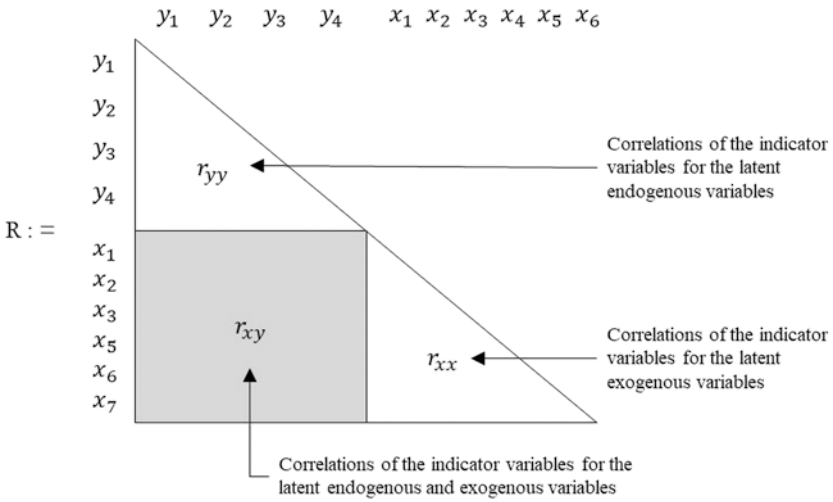


Fig. 2.3 Structure of a correlation matrix. (Source: Own representation based on Backhaus et al. (2016))

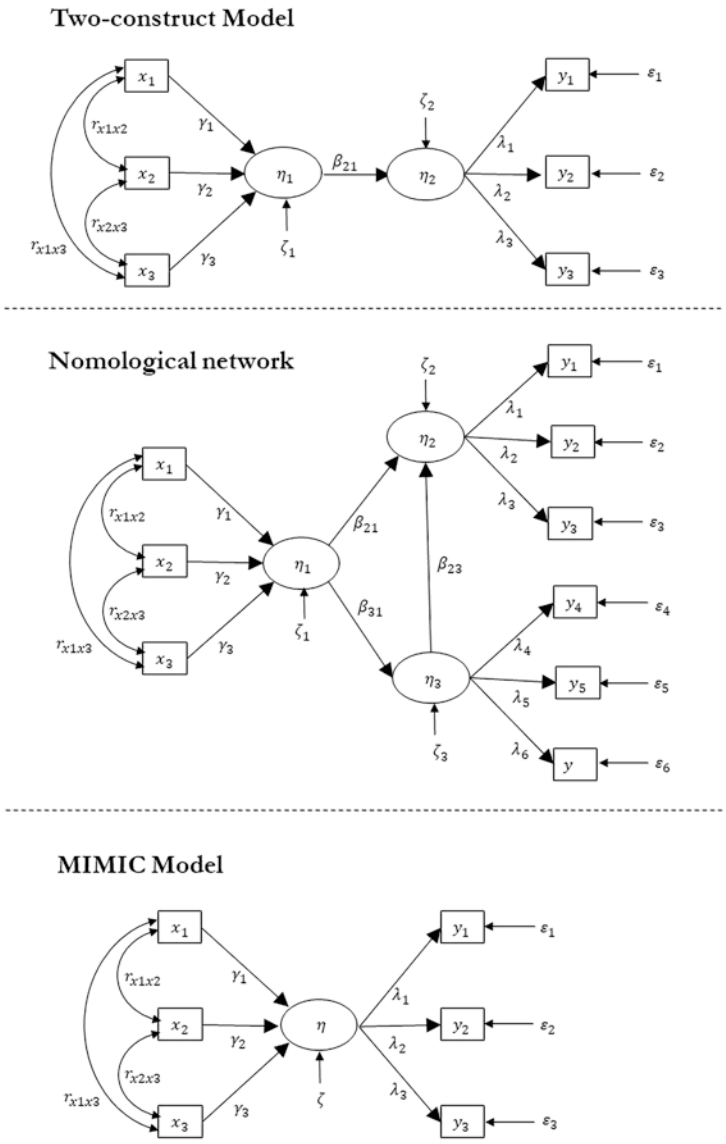


Fig. 2.4 Three models for estimating formative measurement models. MIMIC: Multiple Indicators and Multiple Causes. η: latent variable, ***x:*** formative indicator, ***y:*** reflective indicator, γ: weight, λ: factor loading, ζ: measurement error at latent variable level, ε: measurement error at indicator level, β: path coefficient. (Source: Own illustration based on Diamantopoulos and Winklhofer (2001); Weiber and Mühlhaus (2014)

Two-Construct Model and Nomological Network

In two-construct models and nomological networks (also called multi-construct models), the formative construct η_1 is integrated into a larger model (see Fig. 2.4). The formative latent exogenous variable becomes the predictor of at least one reflective latent endogenous variable. For such a validation, in addition to data for the formative construct, data must also be collected for the reflective construct, or for all other reflective constructs in the case of nomological networks. It must be possible to reasonably assume a theoretical relationship between the latent variables. The path coefficients β_{21} (and β_{31}) represent these relationships; their magnitude, sign, and significance are a necessary condition for construct validity (Diamantopoulos and Winklhofer 2001).

However, integrating a formative measurement model into a larger network does not necessarily solve all difficulties. First, the weights of the independent indicators (y_i) are estimated depending on another construct. If η_2 is replaced by another related construct η_x, a different meaning of the indicators may result. For this reason,η_2 should be as close as possible to the meaning of η_1. If it is then possible to argue that they are even identical and both constructs measure the same thing, the two-construct model is identical to a MIMIC model and the path β_{21} can be read from $\sqrt{R^2}$.

On the other hand, in practice, e.g. when high multicollinearity occurs, it is possible that the model is still underidentified. Theoretically, the model is identified as soon as at least two paths lead from the formative construct to reflective constructs (Diamantopoulos and Winklhofer 2001, p. 271).

MIMIC Model

In a MIMIC model, the construct consists of several reflective and formative indicators ("**M**ultiple **I**ndicators and **M**ultiple **C**auses") and a latent variable (Hauser & Goldberger, 1971, pp. 95-96; Jöreskog & Goldberger, 1975). Reflective indicators are assigned to the formative construct (see Fig. 2.4). Compared to other formative measurement models, a MIMIC model has the advantage that the construct is considered in isolation without including other constructs. Unlike nomological networks, comparability between studies is possible here. Since the quality of the formative construct is tested via the reflective indicators, the same thing is always measured (Weiber and Mühlhaus 2014).

For estimations of formative measurement models using the covariance analysis approach, the programs LISREL ("**LI**near **S**tructural **REL**ationships", Jöreskog and Sörbom 1996), AMOS ("**A**nalysis of **MO**ment **S**tructures", Arbuckle 2012) and Mplus (Muthén and Muthén 2010) are suitable.

2.2.3 SEM in a Variant-Based Approach

Unlike covariance-based approaches, where the covariance matrix forms the basis, a variance-based approach is based on the original data matrix. The difference between observed and model-implied case data (and not their covariances) should be minimized. While in a covariance-based approach all parameters are estimated simultaneously, the variance-based approach consists of two stages. First, the values of the latent variables are estimated from the empirical measured data. These are then used to estimate the parameters. The aim of this approach is to minimize the variance of the error variables in the measurement model ("outer model") and structural model ("inner model") (Weiber and Mühlhaus 2014).

The variance-based approach to estimating structural equation models is two-stage PLS ("Partial Least Squares") modeling. Estimation of the three models described above (MIMIC model, nomological network and two-construct model) is also possible using this approach. However, in order to estimate a formative measurement model in PLS, an integration into such a model is not necessary because formative constructs can also be considered in isolation here. The underlying principle is then a principal component analysis instead of a factor analysis, where the latent construct is understood as the principal component (Weiber and Mühlhaus 2014). In the following, a PLS modeling based on several constructs is described as an example; the differences between LISREL[5] and PLS modeling are discussed at the end of this chapter.

PLS Modeling for Formative Measurement Models

Like any structural equation model, a PLS model consists of an inner structural model and the outer measurement models (see Fig. 2.5). The measurement model in turn consists of the relationships between the latent variables and the indicators.

The structural model is composed of the relationships between the latent variables (η_i):

$$\eta_j = \sum_i \beta_{ji}\eta_i + \zeta_j \quad \text{for all } j = 1,\ldots, J$$

where β_{ji} stands for the path coefficient and ζ_j for the inner residual variable (Boßow-Thies and Panten 2009).

[5] LISREL does not stand for the concrete program here, but for all covariance-based estimation methods in the framework of "linear structural relations".

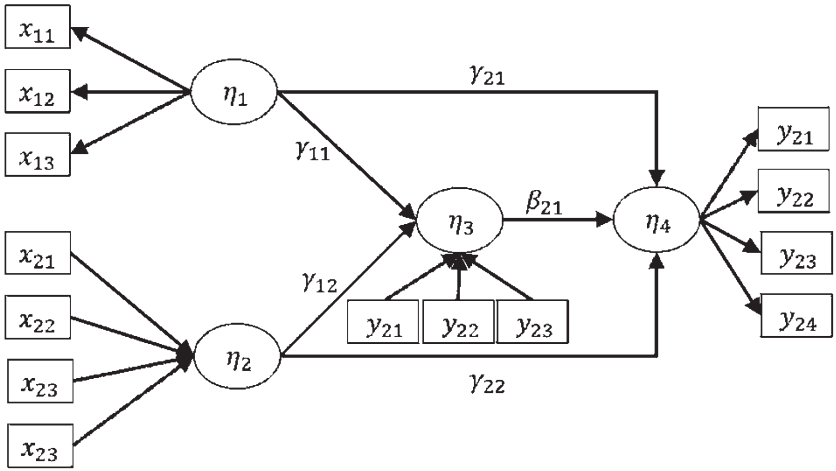


Fig. 2.5 Example of a PLSmodel. (Source: Own representation based on Boßow-Thies and Panten (2009)

In the first step, multiple regressions are used to estimate the construct values (estimated values) for each latent variable for each survey case. The basis for this step is the available empirical information (measurement data). PLS makes use of an iterative estimation algorithm for this purpose (see Fig. 2.6, for a detailed description see Weiber and Mühlhaus 2014). In the second step, the path coefficients (effect sizes) of the structural model are now estimated using a path analysis. Finally, the mean values and constants of the linear regression function are estimated (Weiber and Mühlhaus 2014).

The quality of a PLS model is determined by the explained variance (R^2) of the endogenous variables, which indicates the extent to which the empirical values match the model-implied values of the endogenous variables (Weiber and Mühlhaus 2014). In the meantime, measures of quality such as the SRMR score are also available for PLS models (see example below).

Basically, both approaches presented here are complementary to each other. While the covariance analytic approach is theory-testing (hard modeling), the variance analytic approach is data- and prediction-oriented (soft modeling) (Weiber and Mühlhaus 2014). Some authors believe that for formative measurement models, variance analytic PLS modeling is preferable (cf. e.g. Eberl 2004; Weiber and Mühlhaus 2014). One advantage is that smaller samples can be analyzed with PLS. Furthermore, the better suitability of the PLS method also becomes clear

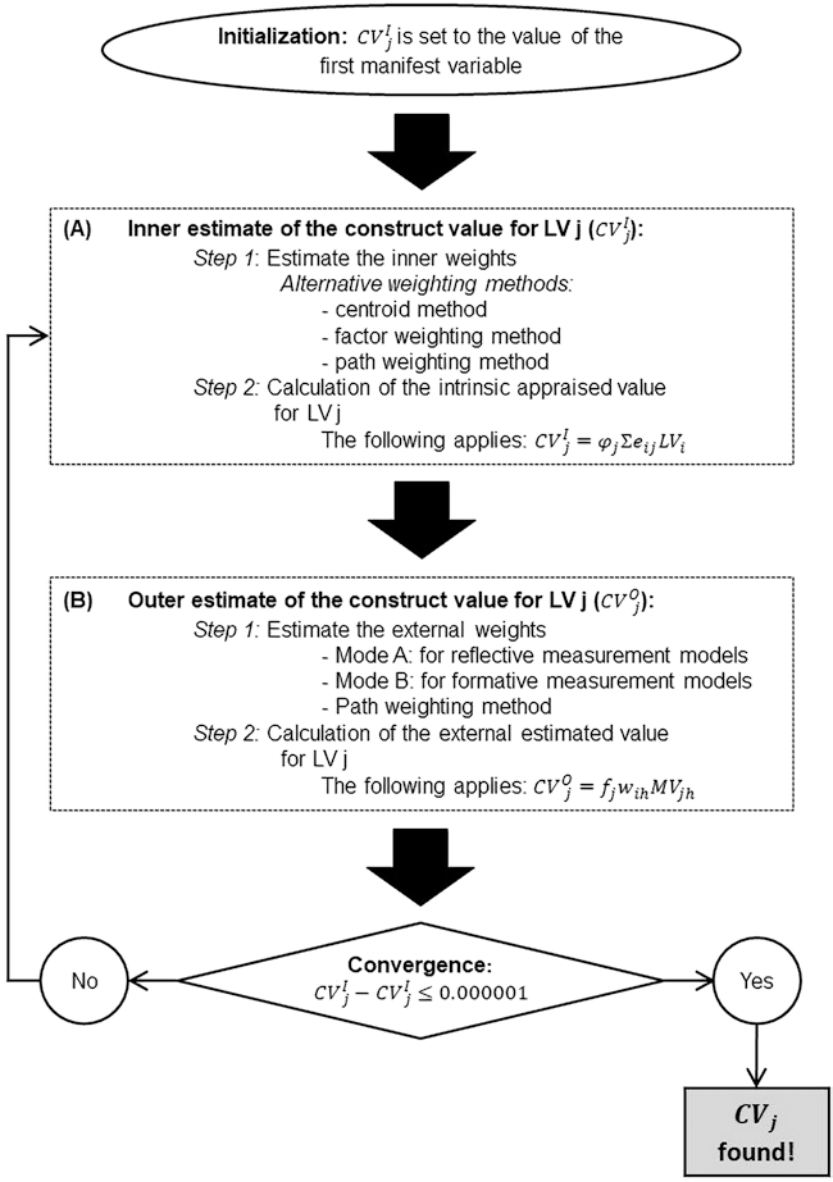


Fig. 2.6 Iterative estimation of construct values. (Figure adapted from Weiber and Mühlhaus 2014, p. 69). (Notes: CV_j^I = construct value of latent variable j (LVj), from the inner model; φ_j = normalization variable for LV; e_{ij} = (inner) weighting variable; LVi = latent variable i influencing LVj; CV_j^A = construct value of latent variable j (LVj), from the outer model; f_j = standardization factor for LVj; w_{ih} = (outer) weighting variable; MV_{jh} = manifest variable j assigned to LVj)

when considering the measurement error. If one understands the measurement error as a combination of cumulative errors of the indicators and not taken into account construct components, then it becomes apparent that covariance analysis approaches are less suitable, since they are concerned with the identification of error variances.

Example

For formative modeling of the construct cyberbullying, Fluck (2020a) specifies a MIMIC model (see Fig. 2.7). Formative indicators are various communication channels through which individuals may experience cyberbullying (referred to as cyberbullying in the figure) (e.g. text messages, chats…). Two other measures of the construct serve as reflective indicators. The first is an aggregate variable (referred to as Σ A1–A4 in the figure) across four forms of cyberbullying, but measured not by the media used, but by the particular type of bullying (insulting, denigrating,…); the second is a single-item question on the extent to which individuals consider themselves victims of cyberbullying (referred to as "rating" in the figure).

Figure 2.7 represents the MIMIC modeling in the variant-based approach, calculated with Mplus.

The model fit, even based on the Chi^2 test, indicates a very good fit of the model to the data. The other indices are also in the good to very good range. The variance in the external criteria is explained by the model to a high degree, in the sum value over the behavior items to 86% and in the rating scale victim experience to 58%. Overall, the amount of variance explained in the construct is 88%. Thus, the model is confirmed.

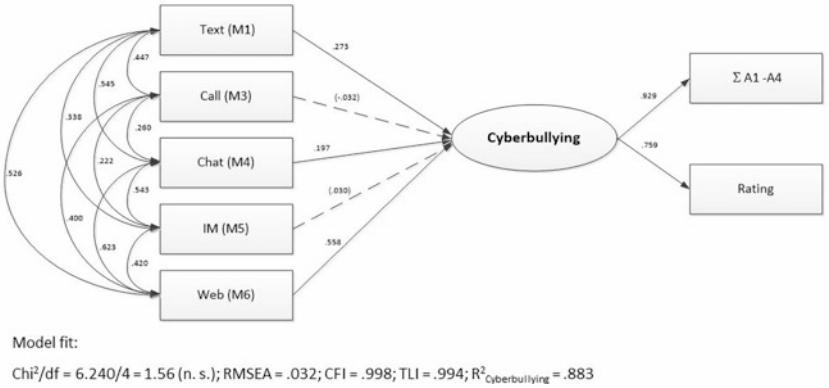


Fig. 2.7 MIMIC model in the variant-based LISREL approach

However, the model also shows that some items must be viewed critically due to the lack of validity as measured by the correlation with the external criteria. Phone calls (M3) and instant messages (M5) are not significantly related to the latent variable. Websites (M6) can be considered the strongest indicator, but text (M1) and chat (M4) also have significant (albeit lower) regression weights.

In order to replicate the MIMIC model from the LISREL estimation in the covariance-based PLS variant as well, two latent variables must be specified instead of just one, since the PLS algorithm does not work with one variable, but always requires at least one endogenous and one exogenous variable. In the PLS variant, the MIMIC model is therefore set up using one formative and one reflective latent variable.

Figure 2.8 shows the comparison of the two variants in schematic representation using an example with four formative and three reflective indicators. Mathematically, the two models are identical (Fornell and Bookstein 1982). The regression parameter, denoted by γ_{PLS} in the figure, can also be taken from the upper model; it corresponds to (the root of) R^2 of η.

Transferred to the present sample data, the representation results from Fig. 2.9.:

In the LISREL[6] model, parameters of the model fit such as RMSEA and CFIT/TLI are based on the agreement between the model-implied (theoretical) covariance matrix and the empirical covariance matrix. The goodness of a PLS model, on the other hand, is assessed on the basis of the explained variance (R^2) of the dependent variable (DV), as this provides information on the extent to which the actual (empirical) values of the DV match the model-implied (predicted by the independent variable(s)) values.[7] Although Esposito Vinzi et al. (2010) furthermore propose a global goodness-of-fit index (GoF),[8] Henseler and Sarstedt (2013) were able to show that this is only usefully applicable in multi-group PLS analyses.

A recently available global statistic for the model fit in PLS estimation is the SRMR (Standardized Root Mean Square Residual) introduced by Henseler et al. (2014), which – analogous to its use in the LISREL model (Hu and Bentler 1998) – is defined as the difference between the model-implied and the empirical correla-

[6] LISREL does not stand for the concrete program here, but for all covariance-based estimation methods in the framework of "linear structural relations".

[7] Due to this fact, the LISREL model is discussed in the literature as a hypothesis testing procedure, while the PLS model is discussed as a prediction procedure (Esposito Vinzi et al. 2010).

[8] The GoF attempts to assess the overall fit of the model by combining the mean of the communalities of all latent variables (as a goodness-of-fit measure for the measurement models) and the mean of the explained variances of all dependent variables (as a goodness-of-fit measure for the structural model) into a geometric mean.

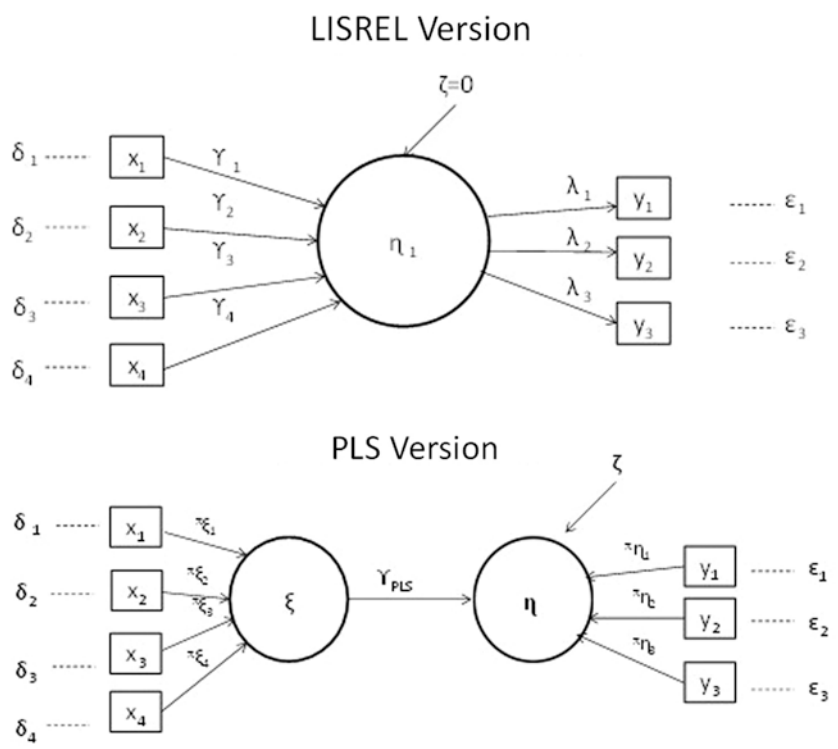


Fig. 2.8 MIMIC model in the LISREL vs. PLS variant. (Figure adapted from Fornell and Bookstein 1982, pp. 445–446)

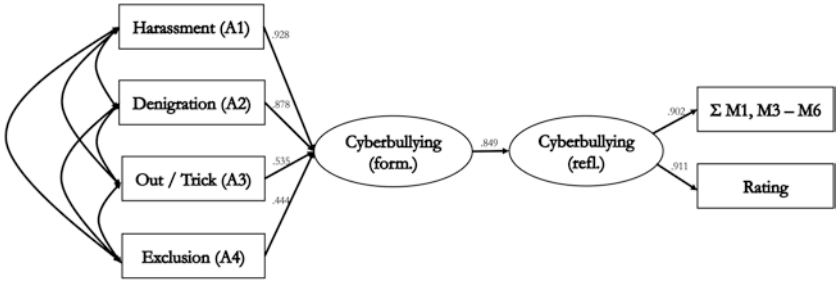


Fig. 2.9 MIMIC model in the variant-based PLS approach

tion matrix calculated from the respective covariance matrices. The smartPLS program (Ringle et al. 2015) in version 3 (3.2.3), which was used to compute the present analyses, outputs two SRMR values by default, one for a *common factor model*, one for a *composite factor model*.[9] Except for models in which only reflective constructs are used, the interpretation of the *composite factor model* SRMR is recommended (Ringle et al. n.d.), which, as is usual for the SRMR, is to be interpreted as a good fit as soon as values <0.10, or values <0.08 in a somewhat more conservative view (Hu and Bentler 1998).

A significance test of the loading parameters is not performed by the program by default, since the PLS method does not assume normally distributed data and the usual parametric significance tests are therefore not automatically applicable. However, smartPLS includes a bootstrapping procedure in which significance tests are performed based on the distributions of loadings, weights, and path coefficients in 500 bootstrap samples. Appropriate calculations show that the regression weight loading parameters are all significantly different from zero ($p < 0.00001$), as expected given the magnitude of the values.

The loadings are all significantly higher than in the LISREL model. However, the literature points out that due to the entirely different logic and algorithms, the results of PLS and LISREL models cannot (and should not) be compared and that considerable differences in the results are to be expected (Weiber and Mühlhaus 2014). The ratios between the indicators are also mapped similarly, but not equally. In both the LISREL and PLS variants, M6 (websites) has the highest loadings, while M3 (calls) and M5 (IM) have the lowest. However, in the LISREL model M1 (text messages) has the second highest loading, while in the PLS variant M4 (chat) holds this position.

The explained variance in the external criteria is also significantly lower than in the LISREL model ($R^2 = 0.883$) with $R^2 = 0.619$. From a purely methodological point of view, the model based on the PLS estimation is also confirmed here with a still good R^2 and a good overall fit ($SRMR = 0.042$).

The modeling based on both approaches is only exemplary here. In practice, users can choose one of the two variants.

[9] For a discussion of the – sometimes varying – understanding of composite factor models, see Bollen and Bauldry (2011).

Formative Measurement Models in Psychology and Education?

3

In this chapter, the question of why increased attention to formative models in psychology and education would be appropriate is explored. To this end, we will first discuss the problems of formative modeling that critics of the method raise against it (Sect. 3.1). Subsequently, the added value of the consistent application of formative models – where they are appropriate – is discussed (Sect. 3.2).

3.1 Critical Aspects of Formative Models

Since the rise in the use of formative measurement models, there has been a lively debate around their basic applicability as well as specific sub-aspects. The authors Hardin et al. (2011) see the cause for the debate and the difficulties of implementing formative measurement models in the fact that advice on how to deal with formative measurement models is mainly based on statistical approaches and not embedded in the theoretical foundation of a test theory.

Implementation

West and Grimm (2014) already see general problems with the implementation of the model, as formative models are difficult to calculate and therefore not very practicable. The approach presented in Chap. 4 shows that this need not be the case.

Markus (2014) postulates that creative researchers could use the concept of formative measurement models to arbitrarily combine all possible indicators into one conceptual unit. In this way, researchers could declare reflective models with

J. Riebel, H. Lichtenberg, *Formative Modelling in Psychology and Educational Science*, https://doi.org/10.1007/978-3-658-39404-2_3

too few intercorrelating indicators to be formative models in *retrospect*, in order to make even poorly fitting models publishable.

Edwards (2011) also criticizes the lack of an internal coherence criterion for formative measurement models – formative indicators may correlate with each other, but they do not have to (see Sect. 1.2). This point forms a basis for some misconceptions among researchers, such as that low correlations can automatically lead to the assumption that a model is formative and not reflective.

The – indeed unscientific – approach put forward by Markus (2014) should be countered by the fact that it is essential to decide on a measurement model in advance, during scale construction, and not a posteriori when the results of a reflective measurement model have not been satisfactory. But the commitment to a specification remains to some extent subjective (see answering guiding questions in Sects. 2.1.1 and 4.1). Therefore, Bainter and Bollen (2015) point out the importance of empirically testing the appropriate specification of the model. Formative measurement models can also be tested for validity (Bainter and Bollen 2015). For this purpose, expert judgements on item selection and also statistical criteria can be used. For this reason, some authors recommend the empirical tests described in Sect. 2.2 (see Bainter and Bollen 2015; Eberl 2004), such as the confirmatory TETRAD test developed by Bollen and Ting (1993, 2000).

Against the argument of unscientificness, it should also be countered that "unscientific procedures" (Christophersen and Grape 2007, p. 105) can sometimes be observed in the reflective model, namely when an artificial increase in reliability is achieved through similarly worded items that merely paraphrase each other. According to Christophersen and Grape (2007), different reflective indicators should also represent different consequences of a latent construct.

Individual modeling variants are also criticized by some authors. Edwards (2011) and Simonetto (2012), for example, find the MIMIC model counterintuitive because the reflective indicators are the consequences of a construct and it is thus inconclusive that these validate the formative indicators (i.e., the causes). Elsewhere (Weiber and Mühlhaus 2014), the MIMIC model is referred to as the standard procedure for formative modeling (at least in a covariance analysis approach). In general, it can be stated that an empirical examination of the appropriate modeling variant is important (Bainter and Bollen 2015). It is up to the user to decide in which way this test should be carried out.

Some authors (see, for example, Edwards and Bagozzi 2000; Howell et al. 2007) also criticize the fact that the formative indicator weights are model-dependent, as there is a need to include other variables in order to be able to estimate the formative indicator weights (Hardin et al. 2011). This in turn has implications for the interpretation of models.

Interpretation

Formative measurement models are mostly multidimensional, as the latent constructs are explained by different heterogeneous indicators that represent different facets of the construct (see Sect. 1.2). However, according to Edwards (2011), multidimensionality is not only a *property of* formative measurement models, but also a *weakness*.

Edwards (2011) postulates that it is not sufficient to have knowledge about the level of indicator-to-construct paths (γ_i) to resolve the ambiguity of the latent construct. This is because, according to Edwards (2011), the variances and covariances of the indicators (x_i) also influence the meaning of the construct.

Another much-discussed problem of interpreting formative models is their susceptibility to "interpretational confounding" (cf. e.g. Hardin et al. 2011; Howell et al. 2007; Kim et al. 2010). Interpretational confounding describes the problem that arises when the empirical meaning of a latent variable is different from that which was assigned to it a priori (Burt 1976). Explained using the classic example of a formative measurement model mentioned earlier, socioeconomic status (SES), this would mean the following: In his original definition, Hauser (1971) described SES using a composite of income, occupation, and housing. Interpretational confounding occurs when a model is calculated in which the construct SES, mainly consists of the function of the variable "income", while occupation and housing have almost no influence (cf. Howell et al. 2007). The different meaning of indicators may influence the conceptual meaning of the formatively measured latent variable, leading to different levels of meaning in different studies and contexts (Hardin et al. 2011). Although Bollen (2007) postulates that interpretational confounding only occurs due to model misspecification, in their study on information systems, Kim et al. (2010) show that interpretational confounding can occur even in correctly specified models. Nonetheless, interpretational confounding is not an exclusively formative problem – misspecification resulting in interpretational confounding can also occur in reflective models (Bollen 2007).

Missing Measurement Error

Edwards (2011) sees another problem with formative models in the absence of measurement error at the item level. The indicators are considered to be error-free and independent (see Sect. 1.2), this assumption Edwards (2011) considers unrealistic as the methods used to obtain these values are subject to error, as is often the case with interviews, observation and/or self-report. However, a formative measurement model also does not claim to explain the variances in the independent indicators. Although the errors in the item values are not quantifiable in the case of independent variables, they are included in the error term ζ together with the unmapped indicators.

Unlike reflective measurement models, construct validity of formative models is more difficult to test because isolated factor analyses are not possible (Edwards 2011). However, the consequences of mis-modeling as a threat to validity apply to both variants. Aspects of valid modeling of formative models are discussed in the following section.

3.2 Formative Modeling of Psychological Constructs?

The above explanations show that formative models are viewed critically by some authors. Despite these criticisms, formative models are part of the repertoire of standard methods in economics. An increased application also in the behavioral sciences is propagated, for example, by Jarvis et al. (2003) in the Journal of Applied Psychology, but so far very little has been implemented.

Psychological constructs, which are also frequently the focus of investigation in educational research, are in most cases typical examples of reflective modeling when they can be viewed as "underlying causes for the performance of certain actions" (Christophersen and Grape 2007, p. 104). In line with the state-trait view of personality (Jäger and Petermann 1995), the concrete behavior and experience (state) in a given situation is substantially co-determined by underlying personality traits that enter into a complex interaction with specifics of the situation.

Nevertheless, there are exceptions that rather suggest formative modeling. Whenever a construct is built by bundling causally pre-ordered individual factors (Albers and Hildebrandt 2006), the reflective notion falls short.

What is applied in economic studies to model the success of companies and management measures (Christophersen and Grape 2007) can equally be transferred to the recording of professional, educational, school and therapeutic success. Conversely, the discontinuation of customer relationships studied with formative models in the marketing field (see e.g. Bruhn et al. 2008) can also be conceptually transferred to educational contexts (e.g. discontinuation of training and studies).

According to Smith (2011), health-related constructs such as social support, coping strategies and emotion regulation require a closer look at the relationship between indicators and construct. If such strategies are initially independent of each other and successful stress coping, emotion regulation, social support, etc. is already given if one of several strategies is used without the use of one strategy also being accompanied by the use of another, the formative specification is obvious.

In fact, there are also a few studies in the field of educational psychology that acknowledge the appropriateness of formative models and implement them accordingly. The following studies should be read by interested readers, as they illustrate the different approaches to formative modeling:

- Konradt et al. (2008) use variance-based (PLS) structural equation models to study learning transfer. In doing so, cognitive and non-cognitive learning strategies are modeled formatively.
- Schmitz et al. (2020) calculate indices from causal factors of teaching competence. The individual factors are included in the overall index to varying degrees, depending on weighting factors determined by experts.
- Ihme and Senkbeil (2017) investigate computer-related competencies of students (ICT literacy). Self-directed PC experiences, instructional support from the family, and instructional support from the school are three constructs that the authors consider formative. The weights in this study result from a complex structural equation model.

The examples show that an application of formative models is particularly useful in the area of competencies and strategies. It is desirable that knowledge of formative models becomes more widespread and that they are applied where indicated.

At the same time, the question arises at this point as to what the consequences of a wrong decision are. How bad is it if we misjudge the "actual" specification type in scientific studies as well as in test and questionnaire construction and reflectively model formative constructs?

While Jarvis et al. (2003) report a significant misestimation of fit values, for Diamantopoulos et al. (2008) the directly resulting methodological problems are manageable. Their simulation studies show that loading parameters are easily misestimated and fit indices change as a result, but not so significantly that models are discarded for lack of fit. Jarvis et al. (2003) estimate the number of formative models incorrectly specified as reflective to be about one-third for the marketing domain. However, in relation to psychological constructs, this proportion is likely to be much lower.

Whether a construct is modeled reflectively or formatively, however, has conceptual consequences in addition to methodological ones. In the reflective "Cronbach's α-LISREL paradigm" (Albers and Hildebrandt 2006) – if applied to formative models – the wrong items are removed. By default, scales are constructed in such a way that many items are included in the analysis that measure the same thing as far as possible. In the worst case, paraphrasing may result in similar items that are hardly distinguishable from one another and thus provide little additional information, although the resulting test length increases reliability (Bühner 2006).

Mis-specified models are not identified as incorrect, particularly in the case of correlating indicators, due to the characteristic values for model goodness and are therefore not discarded. However, it is crucial that the selection of indicators in misspecified models can reduce the validity of the construct (Albers and Hildebrandt 2006). The following example illustrates this issue.

Overview

Professor Sally Smart wants to measure satisfaction with the course she developed, "Reading tea leaves and Bean Counting." The questionnaire she developed captures the construct using seven items that are to be combined into a mean.

Before taking the mean, Sally Smart does a reliability analysis and adjusts the scale. The discriminatory power of five items is satisfactory. These items correlate highly with each other.

- Competence of the teachers
- Accessibility of the teachers
- Motivation of the teachers
- Classroom atmosphere
- Fun during course

However, two other items correlate only weakly with the other items and do not show sufficient discriminatory power.

- Practical relevance
- Fairness of the tests

Sally Smart removes the two items, which are also not highly correlated with each other and thus do not represent a common second factor.

The other five are confirmed in a confirmatory factor analysis on the basis of good fit values. The scale also seems to be valid; at least it correlates significantly with the evaluation forms of the individual courses. The results are also positive: the mean satisfaction score based on the 5 items is 4.2 out of 5. It's just strange that so many students drop out of the course.

Satisfaction is a typical example of a formative construct, provided their causes are captured. Sally Smart was unaware of this fact and, by removing the two low correlating items, increased reliability but trimmed validity. The course of study is fun for the students, but the lack of practical relevance, the arbitrariness of the examinations and perhaps, in addition, poor opportunities on the job market are important facets of "true" satisfaction that are now not captured by the scale and thus severely impair its quality in terms of content validity.

The misspecification leads to the fact that only a part of the construct is measured – but this part is admittedly reliable. In constructing the test, Sally Schlau should have paid more attention to the aspect of content validity, quite independently of reflective or formative modeling.

The phenomenon can also be well illustrated empirically. In a study by Diamantopoulos and Siguaw (2002), the consequences of scale purification according to formative versus reflective procedures were compared by adapting the same scale according to both procedures. In formative scale testing, those items that seemed problematic because of multicollinearity were removed, and in reflective just the opposite, those that did not correlate with the rest of the scale. Thus, of 30 indicators at the beginning of the process, only two (!) remained that were located on both the reflective scale (16 items) and the formative scale (5 items).

The study clearly shows the consequences of the decision regarding the specification type. Validation against central criteria also showed that the formative indicators could explain more variance in the criteria than the reflective ones.

Unlike Jarvis et al. (2003), Albers and Hildebrandt (2006) do not regard the reflective specification of an actually formative construct as an error. It is "not necessarily a false model, [but] a highly restricted model" (p. 13) that focuses on only one or a few facets of an actually multilayered construct. The resulting findings, while not false, are then incomplete, leading to "false conclusions that are ultimately obscured by the application of the sophisticated methodology" (p. 4). The deliberate focus on a particular facet, on the other hand, is legitimate, provided the decision was made consciously and reported transparently.

However, if comprehensive modeling of a formative construct is the more appropriate, researchers should be able to recognize this fact. If this is the case, the test or questionnaire construction or the modeling of the construct must also take into account the specifics of formative models. West and Grimm (2014) consider this to be problematic for standard users: formative models are difficult to calculate and therefore not very practical. In the next chapter, a method will be presented that can be used to calculate formative models in a practicable way and with simple statistical procedures.

Data-Driven Index Formation with the MARI Method

4

The aim of the method developed by Fluck (2020b) and described in the following is to carry out an index construction for formative measurement models that can be realized with standard methods as well as allowing an empirical assessment of the scale quality. This can be done by a *data-driven index construction*, as it is also realized, for example, in the five steps described by Christophersen and Grape (2007) (definition of the construct, determination of the indicators, treatment of multicollinearity, estimation of the measurement model, index calculation) (see Sect. 2.2.1).

In the present variant, index formation begins with mental experiments to formulate a specification hypothesis, whereby the mental experiments can be turned into an empirical question by involving experts (Sect. 4.1). The analysis and (pre-) selection of the items takes place in the second step based on qualitative and quantitative data (Sect. 4.2). The construct is validated using regression analysis methods (Sect. 4.3) on the basis of which the index is formed (Sect. 4.4). Figure 4.1 provides an overview of the procedure. The focus of the following explanations is on the first two steps, as these are considered central and are not described in detail in the previous literature.

In this chapter, the description of the MARI method is illustrated in each case by means of an example from the revision of the questionnaire for recording health behavior (FEG, Dlugosch and Krieger 1995). In each case, the examples are presented in a box and can be skipped over. The following first box provides the theoretical and conceptual background to the case study. Throughout the chapter, central concepts from Chaps. 1, 2 and 3 are repeated, with the focus now on practical implementation. In some cases, several possible methods and approaches are

J. Riebel, H. Lichtenberg, *Formative Modelling in Psychology and Educational Science*, https://doi.org/10.1007/978-3-658-39404-2_4

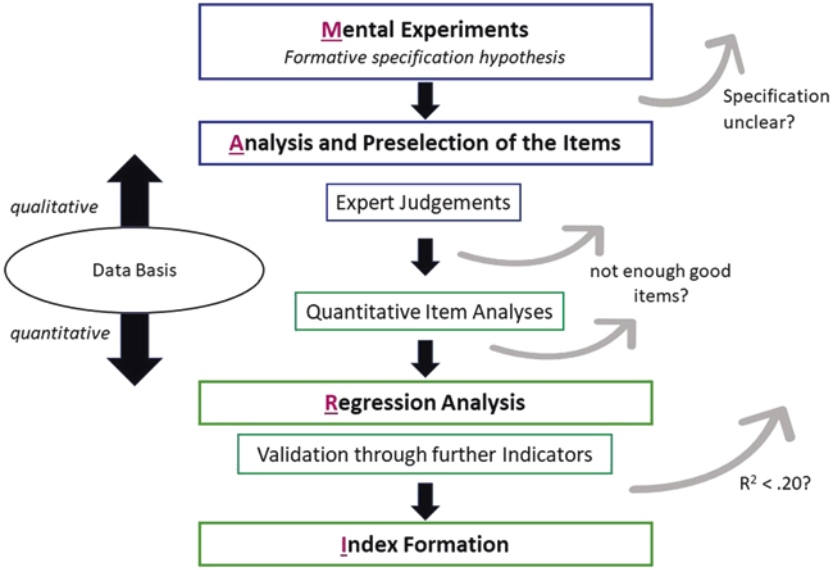


Fig. 4.1 Schematic representation of the MARI method. Grey arrows illustrate the need to stop the process and go back to the conceptual level when the analyses require it

presented, so that readers can select the ones that are suitable and purposeful like choosing from a toolbox.

Revision of the Health Behavior Questionnaire as an Example of the MARI Method

The FEG is a screening instrument for recording various health-related behavioral areas (including diet, exercise, smoking, etc.). Screening instruments are used to obtain an overview of several constructs at the same time and to identify problematic areas. These can then be diagnosed in a more differentiated manner using special test procedures and/or addressed in a diagnostic interview with the test persons.

The FEG is used in health psychological and medical research and diagnostics, e.g. to make comparisons in the health behavior of groups, to evaluate interventions in cure/rehabilitation measures and in the practice of health promotion and counselling.

(continued)

(continued)

In addition to questions on functional links of eating behavior (eating for sociability, to reduce stress, etc.), the nutrition section also contains food scales listing a range of foods that are healthy on the one hand and risky on the other. The scales were developed at the time of the test construction in the early 1990s based on the guidelines of the German Nutrition Society (DGE). Test persons are asked to tick whether they consume a food daily; several times a week; less frequently; or never.

In the course of a currently pending revision of the FEG, the food scales are being revised in order to be adapted to the current state of nutritional science. For example, in the case of bread and rolls, the original 1995 version contains the distinction between wheat (unhealthy) versus rye or wholemeal (healthy), which is now considered outdated, since it is rather the degree of processing of the flour that is important and the use of the whole grain can also be considered beneficial to health in the case of wheat flour, while sifted flours generally contain only few vitamins, minerals and dietary fibres.

The construction of revised food scales is based on theoretical considerations, as well as qualitative and quantitative data along the MARI method.

4.1 Mental Experiments

In Sect. 2.1, the ways of determining whether a construct should be modeled formatively or reflectively have already been described and discussed. Due to the limited informative value of quantitative-empirical *a posteriori* studies, it is recommended that the decision on the type of specification be made *a priori.*

For this purpose, Bollen (1989) proposes the method of mental experiments, which is supplemented by a validation through expert judgements within the framework of the MARI methodology. The mental experiments are conducted along the lines of the questions

- Do the items all measure the same thing, so are they interchangeable?
- What is the direction of causality?
- What are the consequences of change?

A look at existing studies provides initial indications. How has the construct been modeled so far? However, even established modeling variants should be treated with caution, as some models in the literature are likely to be misspecified (Jarvis et al. 2003). It is therefore particularly informative to see work that addresses the issue of formative versus reflective modeling and authors make a conscious decision to choose one type of specification, as Festl (2015) does for the construct of cyberbullying, for example. Unfortunately, however, due to the low profile of formative specification, it is rare for the appropriateness of reflective modeling to be questioned or tested. Rather, in practice, the decision questions are only invoked to justify formative modeling as a deviation from the established standard method. However, it would be important and appropriate, because of the consequences of mis-modeling, to use mental experiments as a matter of principle to justify the specification type, even if one ends up concluding to model reflectively.

The decision-making questions for mental experiments formulated by Jarvis et al. (2003) are implemented in the following presentation in the form of a checklist (based on the decision-making aids in Christophersen and Grape 2007, p. 110). Researchers should first go through this checklist for themselves or in their own team at the beginning of the theoretical examination of a construct or an envisaged type of operationalisation.

In the checklist, some questions are duplicated. The decision support *Can the indicators be understood as an expression of an underlying latent construct? (= reflective)* or *Does the content-related meaning of the construct only result from the interaction of the indicators (= formative)?* is not posed here, for example, as an either-or question, but is implemented in the form of two different yes-no questions. Indeed, it is possible that for one construct both questions have to be answered with "yes". Then the question about the specification cannot be answered unambiguously. In this case, users should not be forced to choose one of the two variants. In this case, only the result of the decision is documented and not the process, during which it may have taken a long time to decide which answer is more appropriate. The implementation in two separate questions allows contradictions and ambiguities to be made transparent.

Readers will notice that the direct question of whether items correlate with each other is not part of the checklist, although empirical approaches such as the TETRAD test answer this very question and even statistically validate it. However, if items do correlate with each other, this confirms the reflective model but does not reject the formative one. Therefore, no information about the correct specification type can be derived from the fact that items correlate. The question about the correlation of indicators was therefore asked as Christophersen and Grape (2007,

p. 110) suggest. They discard the question *"Are indicators expected to be highly correlated with each other in all possible study contexts?"* and instead formulate it as, *"Are constellations conceivable in which indicators are not highly correlated with each other?"* A question that captures the same meaning was implemented in the present checklist (question 9).

No.	Question	Reply		
		Reflective	Inconclusive	Formative
1	Can the indicators be understood as expressions of an underlying latent construct?	Yes	Unclear	No
2	Does the content-related meaning of the construct only result from the interaction of the indicators?	No	Unclear	Yes
3	Are the indicators causes of the construct?	No	Unclear	Yes
4	Are the indicators consequences of the construct?	Yes	Unclear	No
5	In the chronological order, the construct is … the indicators.	Prior to	Unclear	Subsequent to
6	Do the indicators measure the same thing in terms of content[a]?	Yes	Unclear	No
7	Does the content meaning of the construct change if an indicator is omitted?	No	Unclear	Yes
8	Can the items be understood as arbitrary selections from an item universe in which all conceivable items measure the same thing?	Yes	Unclear	No
9	Do the items necessarily have to correlate because of their content importance?	Yes	Unclear	No
10	Does a (superrandom) change in an indicator logically imply a change in the construct?	Yes	Unclear	No
11	Does a (superrandom) change in one indicator logically go hand in hand with changes in the other indicators?	Yes	Unclear	No
12	Is a change in the construct logically accompanied by a change in all indicators at the same time?	Yes	Unclear	No

[a]Note to Question 6: Based on a unidimensional construct – for multidimensional constructs, the question must be posed as "Do all items *within a dimension measure the* same thing?" This consideration applies analogously to other questions for multidimensional constructs

In the interests of scientific honesty, contradictions and contradictory answers must be documented when the checklist is completed and thus made transparent. Ultimately, however, it is neither a matter of a pure "counting" of the answers nor of all the questions on the checklist having to be in favor of one type of specification without exception. Rather, answering the questions helps to become aware of whether a construct is to be modeled (relatively) clearly reflectively or formatively – or whether there is so much ambiguity that no clear assignment can be made. In the latter case, it is worth reconsidering the construct, or the intended mode of operationalization. For example, instead of a formative construct, there may be a reflective one, but one that is multidimensional and thus also covers different facets. However, multidimensional reflective constructs clearly differ from formative constructs in that the multiple dimensions in the reflective model are nevertheless causal *after* the construct. Moreover, more complex cases are often conceivable in which there are second-order factors and the construct is reflective at the first level, but the second-order factor is composed of the various factors in a more formative way (Christophersen and Grape 2007). The reverse is also conceivable.

The formative model is sometimes criticized for the fact that the determination of a type of specification on the basis of mental experiments is based on arbitrary decisions. The decision developed deductively from the theory or based on factual considerations can be influenced by – sometimes even unconscious – expectations of the researchers and is thus necessarily subjective.

This problem can be addressed by validating the mental experiments through the judgment of experts who also conduct these experiments (Albers and Hildebrandt 2006).

Based on the database of expert judgements, a subjective decision is validated (or also rejected) by empirical evidence. In contrast to procedures such as the TETRAD test, however, this empirical decision is made before or within the framework of the test construction.

Since the experts are needed in the second step of the MARI method anyway for scale adjustment, they can also be consulted at the time of the specification decision. For this purpose, they can be presented with the checklist together with the draft items (see Sect. 4.2). The agreement in the judgments of several experts can also be statistically validated with the aid of interrater reliability (Cohen's κ).

The following example shows how mental experiments can look concretely.

Mental Experiments at the FEG

When considering whether the construct "healthy diet" or "high-risk diet" in the FEG should be modeled as formative or reflective, it is worth looking at the items in the original version as a first step, even if these still need to be revised and adapted to the current state of knowledge.

A pre-selection or preliminary version of items is fundamentally important in this first step, so that both the construct and the indicators can be imagined in the context of the mental experiments. The fact that these are supplemented and adapted in the further process is not a hindrance.

Selected items for measuring healthy eating (old version):

- Vegetables (fresh)
- Lettuce
- Potatoes
- Fruit, Fruits
- …

Selected items to measure risky diets (old version):

- Bread / Rolls (Wheat)
- Chocolate, chocolates, sweets
- Fast food (French fries, hamburgers, etc.)
- Sausage

Answering the questions in the checklist now provides information about the possible nature of the constructs "healthy" and "high-risk diet".

No.	Question	Reply		
		Reflexive	Inconclusive	Formative
1	Can the indicators be understood as expressions of an underlying latent construct?	***Yes*** √	Unclear	No
2	Does the content-related meaning of the construct only result from the interaction of the indicators?	No	Unclear	***Yes*** √
3	Are the indicators causes of the construct?	No	***Unclear*** √	Yes

(continued)

(continued)

No.	Question	Reply		
		Reflexive	Inconclusive	Formative
4	Are the indicators consequences of the construct?	Yes	***Unclear*** √	No
5	In the chronological order, the construct is … the indicators.	Prior to	***Unclear*** √	Subsequent to
6	Do the indicators measure the same thing in terms of content*?	Yes	Unclear	***No*** √
7	Does the content meaning of the construct change if an indicator is omitted?	***No*** √	Unclear	Yes
8	Can the items be understood as arbitrary selections from an item universe in which all conceivable items measure the same thing?	Yes	Unclear	***No*** √
9	Do the items necessarily have to correlate because of their content importance?	Yes	Unclear	***No*** √
10	Does a (superrandom) change in an indicator logically imply a change in the construct?	Yes	Unclear	***No*** √
11	Does a (superrandom) change in one indicator logically go hand in hand with changes in the other indicators?	Yes	Unclear	***No*** √
12	Is a change in the construct logically accompanied by a change in all indicators at the same time?	Yes	Unclear	***No*** √

Already in questions 1 and 2 a contradiction arises. On the one hand, the consumption of various foods can be interpreted (question 1) as an expression of an underlying latent construct, for example, if I eat wholemeal bread because I want to eat healthily. At the same time, the construct "healthy diet" does not exist independently of the indicators. This is best illustrated by a counterexample. A person can be intelligent without ever having correctly solved a task on a test. Intelligence may then show itself in his everyday actions, in his success at school and at work, or possibly not at all. Nevertheless, he is an intelligent person. However, healthy eating does not exist independently of actual eating behavior. There can be an "attitude towards healthy eating", in the sense that someone actually finds healthy eating im-

(continued)

(continued)

portant, would like to implement it, but possibly does not manage to do so. However, actual healthy eating is a result of the interaction of many different foods that someone actually consumes. Question 1 thus suggests reflective modeling, question 2 formative modeling.

Once again, the contradictory or slightly unclear character of the construct can be seen in the questions about the direction of causality (questions 3–5). Both directions of causality are both conceivable and theoretically supportable. People may consciously choose and consume foods because they are healthy (and vice versa avoid risky foods). In this case, a healthy lifestyle would be a latent influencing factor behind the actual behavior.

Conversely, however, it could also be argued that people have developed preferences for certain foods because of their learning and life history, and that they like to eat them or simply do so habitually. If these foods are predominantly healthy, the result is a healthy eating style; if they are predominantly unhealthy, the result is a risky eating behavior of which the person may even be aware. At the same time, people may *consider* foods to be healthy, but in fact this is not the case.

Now these causal directions, both of which are conceivable, are not mutually exclusive: They may even apply to different people to different degrees (e.g., depending on the extent to which people care about healthy eating). In the present case, the answer to these questions is ambiguous and not very useful.

Question 7 illustrates once again that "healthy eating" is not a typical formative construct. Omitting an indicator does not change the meaning of the construct. In fact, the construct is *so* complex that a complete list of all possible healthy foods is not even possible. Omitting an indicator probably does not change the meaning of the construct. (Except in the unlikely event that a person's risky eating behavior is based on a single food …).

Nevertheless (question 8), the items are not a random selection from an item universe and thus not arbitrary. The items are not interchangeable. While they all measure healthy or risky eating, they cover different facets of the construct and thus do not measure "the same thing" (question 6).

This can always be well illustrated by imagining that one has to choose one of two items (perhaps for economic reasons) and remove the other from the scale. This is not problematic with a reflective construct. After all, even if the reflective construct is multidimensional, there are multiple items for each dimension, so such a decision is largely arbitrary. In the present case,

(continued)

(continued)
however, where formative modeling seems more appropriate, it makes a big difference whether one asks, for example, about the consumption of high-fat sausage or the consumption of chocolate.

Questions 10 to 12 support the notion of a formative construct. Changes in the construct or in one of the indicators do not have universal consequences. Those who eat more unhealthily (e.g., during a period of stress) will not necessarily increase their consumption of all risky foods. The independence of the indicators is just as clear from these questions as it is from question 9: it is quite possible that the consumption of a healthy food is also accompanied by the consumption of other healthy foods, and the same applies to the unhealthy foods. At the same time, such a correlation cannot necessarily be derived from theory: there is no such thing as *a* healthy or unhealthy diet that is expressed in the same way in all places.

Final Evaluation:

The mental experiments indicate formative modeling. Ticking the category "unclear" is not problematic, as clear statements about e.g., causality cannot be made for all constructs. Ultimately, it is at the discretion of the researchers how they weigh the answers, and there need not necessarily be agreement between experts. Researchers must choose the operationalisation and modeling that they can advocate and justify. It is only important to deal transparently with contradictions and ambiguities.

When dealing with the scales in the FEG, it might be useful to work with two levels: At the first level, items relating to different food groups could be reflectively bundled into factors (e.g., cake and chocolate into one factor, fatty meat and sausage into another), which then form the factor "high-risk diet" at the second level. For this, however, the list of foods would have to be significantly expanded for a more reliable measurement of the individual factors. However, the FEG as a whole is a screening procedure for many different forms of health behavior, which for economic reasons only offers each construct space for a limited number of indicators. Against this background, it seems legitimate to accept the loss of measurement quality associated with a compromise solution and to model the construct at only one level. If this is realized in this way, the mental experiments justify the formative approach.

4.2 Analysis of the Items

Should the mental experiments have led to the conclusion that the formative modeling variant is the correct one, this has consequences for the further course of the questionnaire construction. The usual measures for scale adjustment, as known from classical test theory, cannot be applied due to their covariance-based nature. Own procedures are necessary. First and foremost is the assessment of the quality of the items by experts (Sect. 4.2.1), but there are also statistical analyses in the formative model that are helpful for this purpose (Sect. 4.2.2).

4.2.1 Qualitative Analysis

The assessment of the items by experts should be the basis for answering two central questions.

1. For each item in the questionnaire it has to be decided whether the item "belongs to the construct", i.e. whether an aspect of the construct is operationalized by the item or not. An item may also belong to another, related construct and/or only appear to be related to the construct in question. For example, consumption of low-cholesterol diet margarine is not per se an indicator of healthy eating but is only health-promoting if there is a pre-existing condition or high health risk. Rather, it seems to belong to a different construct, namely adherence to a specific diet. Accordingly, it is assigned to a separate "diet" scale in the FEG.
2. For the construct as a whole, the question of the exhaustiveness of the indicators in their sum must be answered. Do the items cover the construct in its entirety, or are there further aspects/facets of the construct that still need to be operationalized using further indicators? By covering as much of the construct as possible, the "measurement error" of the overall construct should be kept as small as possible.

While the first question can also be answered empirically in the further course of the MARI method, the expert judgments are especially elementary for answering the second question. Items that do not measure the construct in question as suspected are identified, if necessary, by the low correlation with external criteria. However, if not all aspects of the construct are covered by items, this will only manifest itself in a low proportion of total variance explained in the external criterion. However, the low proportion of variance only tells researchers that they have

not sufficiently covered the construct in its entirety, but the numerical value alone does not provide any information about *which* aspects were not considered.

Accordingly, at this point it is a matter of operationalizing a formative construct in such a way that the items measure everything that constitutes the construct. However, if a procedure measures what it is supposed to measure, then the quality criterion of validity is fulfilled, in this case specifically content validity.

"We speak of content validity when a test or a test item actually or sufficiently accurately captures the characteristic to be measured" (Bühner 2006, p. 36). In the reflective measurement model, where items are understood as a random sample from an item universe, content validity describes "how representative the items of a test are of the characteristic being measured" (Schmidt-Atzert and Amelang 2012, p. 145). Even closer to the understanding in the formative measurement model is the definition of content validity that comes into play in the context of criterion-referenced tests. Here, a test is understood to be content valid if it "contains or represents the totality of a set of items" (Klauer 1987, p. 12). Unlike other forms of validity, content validity is not quantified. Thus, it is not empirical validity, but is accepted or rejected based on logical considerations (Bühner 2006). Again, such considerations should not be made by an individual, where they necessarily fall prey to subjectivity, but rather the initial findings of one individual obtained from the literature review are subsequently supplemented by the assessment of experts. The combination of expert interviews and literature analyses should ultimately ensure that "the definitional determination [of the construct] is as broad and at the same time as precise as possible" (Christophersen and Grape 2007, p. 109).

While content validity is the central criterion in the formative model, in the reflective model the quality is rather tried to be proven with forms of empirical validity. At this point, however, it should be noted that content validity is also a central quality criterion in the reflective model, but this is often neglected in practice. Bühner (2006, p. 37) criticizes:

> Test development in particular suffers from the moderate content validity of tests. Tests are often the result of a statistical homogenization process that no longer has anything to do with theoretical foundation. Lack of consideration at the beginning of the design process leads to inadequate procedures already in the development phase.

The procedure for interviewing experts is again best illustrated by the concrete example of the FEG.

After the information provided by the experts has been evaluated, it can be converted into items and the questionnaire can be designed. However, before this is

Consultation of Experts on the Revision of the FEG

Three nutritionists were asked to revise the food scales and were presented with the items in a structured interview. The experts were first asked to state for each item whether it could be retained in the given form or whether it needed to be changed/added to or its language revised. In addition, the experts were asked to state whether the item belonged at all to the construct that the questionnaire designers assumed behind it.

As the following excerpt from the summary of the results shows, interviewee Ms. F., for example, critically noted that "cornflakes, muesli" is incorrectly located on the "healthy diet" scale. However, only sugar-free wholemeal muesli belongs to a healthy diet; sugared muesli in larger quantities is even an indicator of "high-risk diet" due to its high sugar content.

1. Food	Scales	Changes and Additions Ms. F	Scales*
Please tick off how frequently you consume the food items listed here:			
Bread/Rolls: wheat, mixed wheat	R	Bread/Rolls not wholemeal flour	
Bread/Rolls: rye, wholemeal	H	Bread/Rolls wholemeal flour	
Cornflakes, Muesli	H	Whole-grain muesli without added sugars (e.g. with nuts, dried fruits)	G
		Cereal (e.g. cornflakes, cocoa puffs)	Quantity-dependent, tendency to R
Cake, Cookies	R		

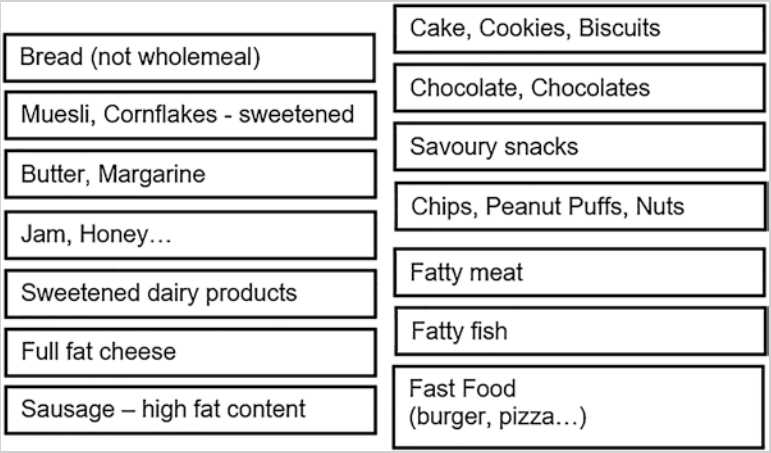
Questionnaire items were then revised if the experts agreed in their judgements. In addition to the items on bread and muesli, the experts also unanimously found the item "salad" to be critical. Test persons could also classify sausage salad, egg salad or pasta salad as salad. One expert suggested renaming the item "raw vegetable salad". However, this suggestion was not implemented because even for raw food salads, the method of preparation has a significant impact on whether it is a healthy meal or a high-risk meal. Due to

(continued)

(continued)
this ambiguity in the possible understanding of the item, this was dispensed with in the revised version.

The second question to the experts relates to the content validity of the procedure. Here, the interview partners were asked to make suggestions for the addition of further items in addition to the evaluation of the existing items. Which essential components of a healthy diet are missing from the original version? Here, the experts suggested, for example, the regular consumption of legumes.

The following overview shows the items for the "high-risk diet" scale after the revision based on the expert assessments. In the following subchapters, this scale will be used again and again, as it will be subjected to a quantitative analysis in the following steps.



used in a more extensive empirical study, the questionnaire should be tested on a few people from the target group in a small qualitative preliminary study. The quality of an item is also essentially based on its comprehensibility before all considerations of measurement accuracy and validity (Christophersen and Grape 2007).

Requirements for an Empirical Study
On the basis of the empirical study, not only the items themselves are to be analysed, but for economic reasons the validity of the questionnaire is also to be tested with the help of regression analysis methods. Thus, the study is to serve as a data basis for the two steps 4.2 and 4.3 in the MARI model. Accordingly, considerations must already be made at this point as to which other constructs, which global questions, and/or which reflectively measured indicators belong to the construct in question and must be operationalized accordingly. If structural equation models are to be used in the course of the regression analysis, this requires a larger sample, the size of which depends on the number of parameters to be estimated but should generally comprise at least N = 100 persons (Backhaus et al. 2003).

4.2.2 Quantitative Analysis

In the quantitative analysis of the items following the empirical study, the focus in this step is on the examination of the items per se. Validation of the items against related constructs or other measures of the same construct is realized in the next step. Even if the usual methods of Classical Test Theory are not fully applied, the items can still be examined with regard to some essential properties.

Descriptive Statistics
The examination of descriptive measures such as mean and standard deviation helps to identify floor and ceiling effects. If items show no or only little variance, they are not suitable to differentiate between different expressions of the latent variable.

The example in Table 2.1 and Fig. 4.2 illustrates a problematic case with the questionnaire for recording cyberbullying experiences, which has already been discussed several times.

The inspection of mean values, standard deviations and distributions of the responses to the respective response categories shows that there are clear ground effects for all items. As is frequently encountered in aggression research, this is a case of zero-inflation, in which one of the extreme categories was selected by almost all subjects (zero-inflated data, Tu and Liu 2002). This poses special challenges for the further analysis of the data, in particular the need to apply nonparametric procedures (Table 4.1).

Fig. 4.2 Distribution of responses to a zero-inflated item

Table 4.1 Example of zero inflation in the data

Item description	M	SD	(1)	(2)	(3)	(4)	(5)
Bullying via text	1.19	0.495	473	74	8	2	2
Bullying via mail	1.01	0.084	554	4	–	–	–
Bullying via phone call	1.14	0.530	507	36	6	3	5
Bullying via chat	1.22	0.591	458	71	12	4	4
Bullying via messenger	1.10	0.445	517	31	–	4	3
Bullying via website	1.33	0.737	426	99	14	9	8

Notes. (1) = "not experienced at all"; (2) = "only once or twice"; (3) = "two to three times a month"; (4) = "about once a week"; (5) = "several times a week"

Even in the case of the item with the greatest variance (bullying by website), the responses are basically only distributed between the first two response alternatives (see Fig. 2.8). The item "Bullying by mail" cannot be "saved" even with non-parametric methods. The practically non-existent variance of $s^2 = 0.007$ makes it necessary to exclude the item from further analyses. At this point, however, it is important to investigate whether the non-existent variance is specific to the sample in question or whether it is also likely to be found in the population. In the present case, the fact that young people do not experience bullying by mail can be explained by the fact that they largely do without e-mail as a communication medium in their everyday lives. This is shown by data from representative surveys of the same year on general media use behavior (mpfs, 2014). Based on this explanation, it therefore seems legitimate to remove the item from the scale.

The example shows that the descriptive statistical analysis of items can provide valuable information about the construct. This also applies to the relationships between the individual items.

Correlation Between the Indicators

As can be seen from the previous explanations, the correlation between the indicators is a much considered aspect of the formative model. A high correlation of the indicators does not automatically mean that it cannot be a formative construct. Also, conversely, a low or non-existent intercorrelation of the indicators does not infer the adequacy of the formative model. Finally, it is also possible that the indicators simply do not measure the same construct – neither formative nor reflective.

Nevertheless, it is worth considering the correlation between the indicators. If only because the reflective model can be rejected if the correlations are not present or are low. Even in multidimensional reflective models, at least a subset of the items must be highly correlated with each other.

Furthermore, procedures such as the TETRAD test can be used to reject the reflective model. The advantage of the TETRAD test (see Sect. 2.1.2) over the consideration of a simple correlation matrix is that the TETRAD test allows a clear yes-no statement about the intercorrelatedness of the items. If one continues with PLS methods in further steps (Sect. 4.3) anyway, the use of smartPLS (Gudergan et al. 2008; Ringle et al. 2015) makes it possible to perform a TETRAD test. Alternatively, the model fit indices in the context of a confirmatory factor analysis can also provide information about the rejection of the reflective model due to low intercorrelation.

However, it must be emphasized at this point that an inspection of the intercorrelations is not obligatory and therefore the calculation of the TETRAD test and CFA can be dispensed with, in particular, if the construct has been specified as clearly formative on the basis of previous steps. Nevertheless, in the following example all mentioned procedures are presented.

Intercorrelation of the Items in the FEG

The analyses reported below were conducted on a sample of N = 281 individuals. The subjects were predominantly students (66.8%) and professionals (23.1%), aged between 18 and 84 (M = 29.24, SD = 11.99). The majority of the respondents were women (18.9% m; 80.1% w; 1.1% d) and in 61.6% normal weight, 5.3% of the sample were underweight, 19.6% slightly overweight, 10.7% significantly overweight (obese).

(continued)

(continued)

An analysis of the risk items in the nutrition section of the FEG shows that there are positive correlations between some items. However, the highest correlation here is only r = 0.485. More interesting is that 31 of 91 correlations are not significant.

The calculation of internal consistency yields a value of $\alpha = 0.745$. Does this not indicate a certain "goodness" of the reflective scale? No. A high value for internal consistency is per se neither a confirmation of scale goodness nor "proof" of unidimensionality or fit of the reflective model. Values such as Cronbach's α are good estimators of reliability *under the condition* that all items measure the same thing, that is, when a unidimensional reflective model is present (Moosbrugger and Kelava 2012). The value does not allow any statements to be made about *whether* such a model is present.

To answer this question, a CFA is necessary. The CFA can be conducted for example in SPSS AMOS (Weiber and Mühlhaus 2014) or in Mplus (Geiser 2011). The loadings of the individual items on the construct range from $\lambda = 0.180$ (chips, flips, nuts) to $\lambda = 0.614$ (sweetened dairy products). The overall model must be discarded due to the indices on model fit (CFI = 0.641; TLI = 0.576; RMSEA = 0.102).

The TETRAD test (performed in smartPLS version 3.2.8 for details on the procedure see Gudergan et al. 2008) confirms this picture. For 14 items there are 1001 possible tetrads, of which 78 are non-redundant. According to the TETRAD test, 15 of these TETRADs are significantly different from 0. The TETRAD test already rejects the reflective model with only one tetrad significantly different from 0.

What can be deduced from these results? It should be noted once again that in practice the combination of TETRAD test and CFA is redundant. In the present case, it is helpful to use at least one of the two procedures. Readers will recall that the decision questions to determine the specification type in this questionnaire were not unanimously answered in favor of the formative model. Even if the formative model is not confirmed by the application of the procedures described here, at least the reflective model is rejected.

The previously assumed assumption that the items do not all measure "the same thing", that they are not interchangeable but operationalize distinct facets of unhealthy eating, could be confirmed here and justifies the further treatment of the construct as formative.

Whether users actually use CFA or TETRAD testing to reject the reflective model, or leave it to inspect the correlation matrix, is at their own discretion.

Formative indicators do not have to correlate, but they may correlate. However, a high correlation between indicators can become problematic if there are strong linear dependencies between the variables and multicollinearity is present.

Multicollinearity

One of the measures mentioned in the literature for formative scale adjustment is collinearity testing (Eberl 2004). If items represent nearly perfect linear combinations of each other, both content-related and technical difficulties arise (see Sect. 2.2). Several procedures exist for detecting multicollinearity. Purely based on the inspection of a correlation matrix, multicollinearity cannot necessarily be detected, since multicollinearity can also arise in the case of low bivariate correlations due to the combined effect of several variables (Schneider 2009).

Pragmatically, the variance inflation factor (VIF, Backhaus et al. 2003) is used because it can be easily calculated. With the VIF, each indicator is regressed on the other indicators by means of multiple regression and with

$$VIF = \frac{1}{1 - R^2}$$

a measure of multicollinearity is calculated. A VIF > 10 (corresponding to an R^2 of 0.9) indicates particularly problematic items, however, values of VIF > 3 are already considered critical by some authors (Diamantopoulos and Riefler 2008; Weiber and Mühlhaus 2014). At VIF < 2, items can be classified as unproblematic in any case (Schneider 2009).

Diamantopoulos and Winklhofer (2001) advocate removing multicollinear items from the formative scale since the construct in question is already sufficiently covered by the other indicators. This "easy way out" solves the technical problems posed by collinear items in subsequent regression analytic procedures. However, Christophersen and Grape (2007) argue against such elimination on statistical grounds, arguing that this would result in a loss of information, albeit a small one. To deal with this problem, it is recommended to combine the multicollinear indicators directly into an index (Sect. 4.4) and to continue working with this index (Albers and Hildebrandt 2006). The advantages and disadvantages of both approaches are discussed in the following box.

Multicollinearity in Practice: Study Satisfaction

In the FEG data example, with 1.19 < VIF < 1.70, there is no multicollinearity in the indicators. Another example will therefore serve to illustrate the problem.

(continued)

(continued)

When evaluating a study program, a number of indicators are used to survey various aspects of student satisfaction (Funke et al. 2010). In detail, it is asked how satisfied the students are with …

- the study plan
- the work-life balance
- the contents of the courses in general
- the professional qualification through the study program
- the expected career opportunities after graduation
- the general information about the elective modules
- the counselling by the academic counselling service
- the theoretical and practical relevance in the design of studies
- the accessibility of the lecturers
- the professional competence of the lecturers
- the forms of examination
- the level of difficulty of the tests
- the premises
- the information on the institute's homepage
- the teaching materials available
- the selection of literature in the institute's library

With $1.905 < VIF < 11.651$ and only 4 items with $VIF < 3$, almost all items have a critical value. What is to be done?

First alternative: item selection

Since almost all items show an increased VIF, a selection of indicators here naturally only makes sense if one applies a somewhat strict criterion of $VIF > 10$. Then there are two items that have such a value greater than 10: *Teaching Materials* and *Accessibility of Lecturers*. One could now remove these two items and subject the rest of the indicators to external validation as described in Sect. 4.3. In such a validation, the question "*How satisfied are you with your studies in general?*" would be used as a further measure of satisfaction and regressed to the individual indicators.

But is it at all legitimate to do without these two variables? The lecturers are also the subject of evaluation by students in a second variable, so that the question can be raised as to whether the lecturers of the study program must be evaluated with two items. However, this second item is not about accessibility, but about lecturers' competence. Both items correlate to $r = 0.599$ ($p = 0.01$) with each other. A significant proportion of the variance in the

(continued)

(continued)
problematic item accessibility *of lecturers* is thus covered by the item on their competence. Nevertheless, it can be argued that competence and accessibility of lecturers are different aspects that do not conform in all cases.

The problem becomes even clearer with the item *teaching materials.* If this indicator is omitted, the corresponding aspect of satisfaction is simply not included. The VIF of the item is 11.651 – thus 91.42% of its variance can be explained by the combination of the other indicators. Nevertheless, there may be reasons for not wanting to do without the item. The indicator may make a contribution, however small, to incremental validity, but it *is* a contribution – as argued by Christophersen and Grape (2007), who argue against removing indicators.

Second alternative: index formation

Forming an index directly from the indicators (see Sect. 4.4) has the great advantage that no indicators need to be omitted that might define essential facets of the construct.

However, in doing so, one foregoes potentially valuable information that could be provided by the regression analyses. Even with collinear items, it is therefore recommended to carry out the third step (Sect. 4.3), although special caution is then required when interpreting the models.

Quantitative Auxiliary Analyses for Content Validity

As mentioned earlier, content validity is a critical moment in formative modeling. The test for account-valid items can also be supported by quantitative auxiliary analyses. Once again, the example of satisfaction is used for this purpose:

The facets of student satisfaction mentioned in the previous box are derived from literature analyses and surveys of experts (Funke et al. 2020). On the one hand, it can be assumed that some of these factors influencing satisfaction are "closer to the construct" than others, i.e., that, for example, premises in terms of size, equipment and accessibility are certainly important for students, but that, for example, the competence of the lecturers is likely to have a greater influence. On the other hand, students are likely to differ in terms of which satisfaction factors are decisive for them. Validation based on external criteria (e.g., in the present case by an overall measure of satisfaction) can answer the question of the quantifiability of the relative influence of individual indicators (see the following section).

Alternatively, however, if the scope of a survey permits it, content validation can also be carried out at the individual level. In the study by Funke et al. (2020), for example, the question was not only asked for each facet "How satisfied are you

with …", but also "How important is the following factor for you to be generally satisfied in your studies?"

The comparison of the mean values shows which indicators are of great importance and which are of lesser importance. For all items, the mean value[1] lies between 1.04 (lecturers' competence) and 2.58 (premises). This means that all values are above category 3, which is headed "not very important". All in all, this suggests that the items should be retained for the time being and investigated further. At the same time, the importance directly ascertained here can be an alternative weighting factor for the index formation (see Sect. 4.4).

A look at the standard deviations also provides information on how unanimous the respondents are in their assessment of the importance of an influencing factor. In the present example, the professional competence of the lecturers, which was described as important or somewhat important by all respondents (M = 1.04) with SD = 0.204, is a universally central characteristic. The importance of the literature selection in the institute library, on the other hand, is a clearly more "controversial" feature with SD = 0.900, although it is also rated as important on average (M = 1.87). This unanimity of the respondents, which can be read from the dispersion of the answers, can be important in the further course of the decision-making process regarding the item selection. If it is known that a facet may generally have a comparatively low influence on the target construct, it can still be retained if a high standard deviation indicates that it appears to be central to at least some of the respondents.

4.3 Regression Analytical Validation

When testing the quality of scales within the framework of the paradigm of classical test theory, items are usually selected in a first step in order to ensure high reliability. In this process, those items are removed that do not correlate highly with the other items and thus with the rest of the scale.

In a second step, the construct is validated by correlating it with other constructs and with equal measures of the same construct. The correlation should be in accordance with the theory in order to ensure the validity of the scale.

The logic of this approach is based on the assumption that all items measure the same construct but do so differently. Items with low discriminatory power are therefore not reliable. However, this logic does not apply (meaningfully) to forma-

[1] The scaling included the categories 1 = very important, 2 = somewhat important, 3 = not very important; 4 = not important at all.

tive constructs. Items with low discriminatory power may well measure the construct reliably – but they may measure a different aspect of the construct.[2]

Validation must therefore already take place at the level of individual items and not only when the construct has already been formed. On the contrary, the decision whether an item belongs to the construct or not is also dependent on the validation. In the following, a procedure is presented that supports such a decision. In the formative variant, too, there is a two-step procedure that first examines the individual indicators before subsequently validating the construct as a whole. However, unlike the reflective variant, other measures of the construct (or similar constructs) come into play for validation at both points. It should be noted that, as in other places in formative modeling, the credo applies that numerical results do not entail dogmatic guidelines for action, but together with content-related considerations and interpretations of the statistical data only form the basis for informed consideration.

In both modeling approaches, the first step in the quality test is an analysis of each individual item. In the reflective model, the individual item is compared with the other items; the highest possible common variance is desirable and is considered an expression of reliability. In the formative model, the individual item is examined for shared variance with an external criterion or other measure of the construct. The second step examines the items in their interaction to determine how much variance they can jointly explain in a criterion.

4.3.1 On the Choice of Regression as a Method of Analysis

Albers and Hildebrandt (2006) make the suggestion "that one should revert to regression analyses in the case of exclusively formative indicators" (p. 2). What do the authors mean by "revert"? In fact, it is a matter of abandoning a frequently taken path that offers some advantages, but in the present case can be replaced by a kind of shortcut.

Construct validity is tested in reflective measurement models with structural equation models, i.e., with a latent modeled regression analysis,[3] which are combined with confirmatory factor analyses to define the latent variables. Compared to manifest modeling, these CFAs provide the added value that information on the

[2] As Albers and Hildebrandt (2006) point out, a false impression of one-dimensionality may also arise in reflective models if other facets are removed in the first step before the link to validating criteria is established in the second step.

[3] Or in the case of a mesh of several variables with latently modeled path analyses.

reliability of the individual indicators can be taken from the model and that the estimated proportion of the measurement error is separated from the true value.

Such a separation into measurement error and variance explained by the model is of less value in formative models because the indicators are specified here as *independent* variables and their measurement error remains unknown in detail. Instead of asking how much variance the construct explains in the items, the formative model asks how much variance in the construct can be explained by the items. Modeling at the latent level is therefore still possible for formative models (see Sect. 2.2) but not absolutely necessary (Albers and Hildebrandt 2006).

In doing so, we dispense with the fit indices issued for structural equation models, which allow a statement to be made about how well the selected model describes the structures actually present in the data. However, other measures are more important anyway: the validation of the individual formative indicators on the one hand (1) and the entire list of indicators on the other hand (2) is centrally concerned with the question that was also illuminated in step 2 of the MARI method (Sect. 4.2) in the context of the expert interviews.

1. Does an item actually measure an aspect of the construct?
2. Is the construct sufficiently described by the interaction of the indicators or are essential aspects missing that have not yet been operationalized?

Both questions can be made more precise:

1. Is the item significantly related to the construct? Is there a significant β-weight in a simple linear regression?

   It should be noted that a significant correlation between item and construct is only a necessary, but not a sufficient condition for the item to belong to the construct. It could also be an expression of another construct that correlates with the one examined here. Thus, the statistical version of the question can ultimately only falsify the assumptions. Against this background, the need for an interpretation of statistical results that is always coupled to content and logic is reinforced: "When deciding on elimination, it must always be weighed up whether the removal of an indicator can be considered justified from a theoretical point of view" (Christophersen and Grape 2007, p. 113).

2. Can the items collectively in a multiple linear regression account for sufficient variance in the criterion?

Here, the focus is on the R^2 in the regression model. While the significance level and effect size of the regression weights can be determined according to common conventions in question (1), users are much more challenged in question (2) to make a decision regarding whether and when *sufficient* variance is explained. How large does R^2 have to be in order to say that the indicators in their sum are sufficiently suitable to capture the construct to a large extent?

The answer depends in no small part on the choice of criterion. Validation against a related construct, in contrast to validation against a second measure of the same construct, suggests less high expectations for the proportion of variance explained. Of course, effect size measures can also provide supportive heuristic decision support. For example, for formative models specified with PLS models, a variance resolution of 20% in relevant external criteria has been suggested (Chin 1998; Lohmüller 1989). If the regression analysis for the overall model falls below the guideline value of $R^2 = 0.20$, users should question whether (a) the dependent variable was close enough to the construct or (b) the indicators sufficiently represent the breadth of the construct. In the following, the procedure is illustrated using the example of the FEG.

Regression Analytical Validation at the FEG

In the first step, the indicators of high-risk diets are examined individually. Readers will recall that the selection of questions was based on theory and interviews with experts. A global question on healthy eating is used as a validating criterion. The question is: "I eat a healthy diet, i.e. varied, wholesome, low-fat, etc.".

The following table shows for each item the proportion of explained variance on the criterion as well as the regression weight.

Item	R^2	β	p<
Bread/rolls – no wholemeal	0.055	−0.234	0.001
Sweetened muesli/cornflakes	0.027	−0.164	0.007
Butter	0.048	−0.218	0.001
Sweetened dairy products	0.066	−0.257	0.001
High fat sausage	0.084	−0.289	0.001
Cake, cookies	0.031	−0.175	0.004
Chocolate	0.073	−0.271	0.001
High fat meat	0.083	−0.287	0.001
Fast food	0.103	−0.322	0.001

Note: Items that are crossed out do not have significant weights

(continued)

(continued)

It is noticeable that some indicators are not significantly related to the criterion and others show significant effects. In the present case, the food items "bread (not wholemeal), butter, sweetened dairy products, sausage, chocolate, meat and fast food" appear to be particularly influential. The items "jam, high-fat cheese, salty snacks, chips and high-fat fish", on the other hand, need to be checked. If the validating criterion is a very good measure of the construct in focus, then consideration can be given to removing these items. *Consideration* means that items must not be deleted blindly, but that a discussion with experts, a look at the literature and/or factual considerations must be taken into account at this point.

Caution is called for in interpreting the data and, in particular, in drawing conclusions from them. In this case, it is important to remember that the items are not objective measures of actual food consumption, but rather subjective self-assessments. Also, the criterion does not measure whether a person eats a healthy diet! It only provides information about the extent to which a person *is convinced that they* are eating healthily. It would therefore be wrong to conclude that foods that have not become significant are not risky. The only conclusion that can be drawn is that an increased consumption of these foods tends not to be perceived as unhealthy by the test persons. The items that are not significant in the analysis should therefore not be removed from the list without consideration. In the present example, it is conceivable that subjects are not aware of whether a food is risky. Instead of a purely number-based selection of items, it therefore makes sense to re-examine the content of the items. For example, some respondents might find it difficult to distinguish spontaneously between low-fat and high-fat cheese, which could explain the poor performance of the item.

Similar to the procedure in a Delphi study (Häder 2009), it can be a sensible step at this point to present the empirical results to the experts again and to interpret them together before items are changed or even removed.

In the second step, the item list is validated as a whole. First of all, a multiple linear regression with the criterion "healthy diet according to self-assessment" already used for the individual analyses yields an explained variance ratio of $R^2 = 0.216$. Considering both the distortions due to self-assessment and the fact that only a small selection of foods is asked for in this screening procedure, this value can be regarded as satisfactory.

(continued)

(continued)

In the multiple regression, however, only the items "butter, chocolate and fast food" now have significant β-weights. Betas are to be interpreted as partial correlations, i.e., they measure the influence on the criterion adjusted for the influences of all other variables in the model. If a β-weight is not significant, the item does not add any variance explanation beyond the influence of the other variables. Nevertheless, in the formative model, one should not remove the nonsignificant items if they had significant weights in the individual analyses. In the formative model, correlations between items are possible and often occur in practice. However, they do not necessarily have to occur; the correlation is not derived from a theory.

If a reflective model were available, the item "sausage" could be removed, since it does not explain any variance beyond "butter, chocolate, and fast food." To put it bluntly, this would mean: Those who eat butter also tend to put a slice of sausage on their bread, so we no longer need to ask about sausage specifically. This approach is of course nonsense! If the item "sausage" were omitted, the unhealthy dietary behavior of a person who eats a pound of sausage every day but replaces the butter with cream cheese would not be noticed. This example clearly shows the necessity of not only referring to the overall model when selecting indicators, but also to consider the individual analyses.

4.3.2 Selection of Dependent Variables

Regression analyses require a dependent variable to be explained. In contrast to the reliability analyses in the reflective model, the assessment of item quality is not based solely on the relationship of the items to one another. What is needed is another measure of the construct in question that can be related to the formative indicators. The crucial question is obvious: What is a good measure for the construct? Given that there would be little need to develop a new measure if *the* ideal measure already existed, this question is fundamentally non-trivial (Rossiter 2002). Validation using similar but distinct constructs (see, e.g., in Christophersen and Grape 2007) is also possible but does not solve the problem of selection. Different variants are conceivable, and ideally several are implemented:

1. The validating variable can be a global measure that captures the construct as a whole. The example of "healthy eating" shown in the box above already indicated possible limitations of the global question. Users need to be aware that global questions and multi-item lists do not necessarily measure the same construct. Often, the global question is subject to the disruptiveinfluence of respondents' subjective perceptions even more than individual items. Fluck (2020a) discusses this issue using the construct "cyberbullying" as an example. While multi-item lists specifically ask about the experience of certain forms of virtual violence and thus capture a countable quantity of incidents, a global question ("To what extent do you see yourself as a victim of cyberbullying?") transfers part of the operationalization to the respondents. By prefacing the question with a description of the construct, the aim is to create a mental representation that respondents are asked to compare to their own situation. If such a global question is used as a validation for the aforementioned multi-item list, it must be taken into account when interpreting the variance explanation that the same construct is not being measured: "self-perception as a victim" and "number of assaults experienced" overlap but are not identical. Thus, as empirical studies show (e.g., Hamby and Finkelhor 2000), there are always people who describe themselves as victims even though they have experienced hardly any assaults and, at the same time, others who have no self-perception as victims despite a factually large number of victimizations.

   The problems described do not apply to all global issues, but they must be anticipated.
2. The use of other measures for the construct in question, which are not asked in a global question but based on several items, is also suitable for the validation of individual indicators. These can, in turn, be calculated into an index for use in a regression, or – if they are reflective – they can be linked to the formative items using a MIMIC model. Some authors champion the MIMIC model as the standard for testing formative measurement models (Weiber and Mühlhaus 2014). However, MIMIC models require structural equation modeling and are not implementable with simple regression analyses, even though they conceptually represent latent-level regression. They also require the ability to measure a construct both reflectively and formatively. Then, the reflective scale can be used to validate the formative one, thereby not examining criterion validity, but rather the construct per se – "from within" (Craven et al. 2013, p. 68). This can be particularly useful with little-studied constructs if a conceptualization independent of other constructs is to be undertaken (for an implementation using the example of the construct cyberbullying, see Fluck 2020a).

Albers and Hildebrandt (2006) criticize the MIMIC model for eliminating indicators based on statistical validation, which has a particularly negative effect when the reflective indicators capture a highly restricted area of the construct. Therefore, it is important that the reflective indicators are able to capture the construct in its entirety and do not only refer to partial aspects.

3. At the manifest level, validity can also be examined through the relationship with other constructs. Criterion validity can be interpreted prognostically if, for example, the indicators are supposed to predict the dropout from an education, and it is examined to what extent such a dropout can be predicted by the interaction of the indicators.

Likewise, the connection with other constructs can validate the formative items. The expected effects should then be estimated to be correspondingly lower than is the case when validation is based on measures of the same construct. In the example of the FEG, the explained proportion of variance when using the global measure "healthy diet" showed a medium effect with $R^2 = 0.216$ (see above). If the question about "conscious nutrition" is used as the dependent variable instead, the value falls to $R^2 = 0.148$. The explained proportion of variance becomes even smaller if the food consumed is related to the body mass index BMI; it is then $R^2 = 0.111$. If one considers the many other possible factors influencing BMI (genetics, exercise behavior, occupational activity, age, illnesses, etc.), however, even this small effect is highly significant.

4.4 Data-Driven Index Formation

In indexing, the formative indicators are combined into a single variable. This variable can be used for further analyses. For individual case diagnostics, the distribution of the index variable can also be a reference point for classifying results. This is discussed below with an example.

For the index formation, the results from the regression analyses in step 4.3 can be relevant as a basis for decisions regarding the item selection. A more detailed description of different possibilities for index formation has already been given in Sect. 2.2.1.

In the simplest case, the index can be calculated using a sum value or a mean value across the indicators. As a rule, a mean value is preferable for personality tests, since on the one hand it compensates for the problem of individual missing values and on the other hand it can be interpreted in the metric of the original re-

sponse categories.[4] If not all items are based on the same response scale, the data must be standardized.

Two decisions have to be made in the fourth step: First, either all indicators can be netted, or a selection can be made. Second, can the (possibly remaining) indicators all be weighted equally, or should they be given an individual weight (Albers and Hildebrandt 2006).

The first decision depends on the state of conceptualization of the construct. Is it known from the literature what the key facets are? Do established measurement procedures exist so that all indicators are demonstrably relevant? If not, it is obvious – in the case of newly developed item lists – to check whether the indicators actually measure partial aspects that belong to the construct.

It has already been pointed out in the previous subsection that item selection should not be done lightly. One rationale may be to calculate regressions on several different relevant constructs and then retain those indicators that are significantly related to at least one of the constructs. Ideally, the decision is validated communicatively with experts. In conversation, the statistical results can be interpreted together.

The second decision concerns the relative influence of the individual indicators. Weighted sum/average values are particularly indicated if (a) the use of weights in general and (b) the specific selection of the individual weights can be well justified.

If a measure very close to the construct existed for the regression in 4.3 as the dependent variable for validation, the respective regression weights can be used as weighting factors in index formation (Albers and Hildebrandt 2006). Modeling in structural equation models also results in regression weights that can be used as weighting factors for index formation (for an example, see Ihme and Senkbeil 2017).

In addition to these quantitative data from the regression analyses, other sources can also be used as a basis for weighting (Christophersen and Grape 2007). In addition to the use of weights based on the literature (e.g., by adopting weighting factors from previous studies), the judgements of experts can also be used here. These can determine which items are particularly important and should therefore be included more strongly in the calculation of the total value (for a practical example, see Schmitz et al. 2020).

In Sect. 4.2.2, under the heading "Quantitative auxiliary analyses for content validity", a procedure was also described in which the weighting factor can be

[4] See Homburg, Hoyer and Fassnacht (2002) for the use of the geometric mean for non-compensatory indicators.

empirically ascertained. In this way, it is possible not only to record the *expression of* a characteristic in a survey, but also to view the respondents as experts of their own lifeworld and to have them give an assessment of how important they consider each individual facet to be for the overall construct. If, for example, the focus is on satisfaction, this describes a sensible procedure.

Individual weighting by the participants can also be useful for individual case studies, for example, if satisfaction with a measure is asked and the participants thus have the opportunity to give special weight to the aspects that are particularly important to them.

In the following, the effect of different weightings is illustrated using the example of the FEG.

Index Formation with and Without Weightings for the FEG

For the sake of simplicity, we assume for the example that the risky diet scale has been reduced to 9 items. Regression analyses yield β-weightings for the items, which indicate how strongly the indicator is related to the construct. These are entered in column 2.

Two (fictitious) persons Klaus and Bärbel fill out the questionnaire. Despite different answer patterns, both receive a sum value of 23. Bärbel likes to eat sweets, whereas for Klaus rather fatty sausage, fatty meat and fast food are problematic (see column "raw value" for both persons).

		Klaus		Bärbel	
Item	β	Raw value	Weight β	Raw value	Weight β
Bread	0.23	4	0.936	2	0.468
Muesli	0.16	1	0.164	3	0.492
Butter	0.22	3	0.654	3	0.654
*Dairy prod.	0.26	1	0.257	4	1.028
*Sausage	0.29	4	1.156	1	0.289
Cake	0.18	1	0.175	4	0.7
Chocolate	0.27	1	0.271	4	1.084
Meat	0.29	4	1.148	1	0.287
*Fast food	0.32	4	1.288	1	0.322
Total		23	6.05	23	5.32

(continued)

(continued)

The value of 23 can be interpreted more meaningfully if it is converted into a mean value. This is 2.56. On average, the subjects consume the various high-risk foods between "less frequently [than several times a week]" and "several times a week". The comparison with the characteristic values of the sample (M = 18.40, SD = 3.77) shows that both Klaus and Bärbel are more than one standard deviation above the mean. A large part of the sample [PR(85) = 22] eats healthier than Klaus and Bärbel.

But are the eating habits of Klaus and Bärbel really equally bad? Bärbel eats a lot of sweets and sweet dairy products, but for Klaus it seems to be the main meals, which consist of fatty meat and fast food every day. According to her own statements, Bärbel does not eat either of these at all.

If the index for unhealthy nutrition is not formed based on a simple sum value, but on the basis of a weighted sum value, a different picture actually emerges.

The β-weights from the regression analysis give each item a specific weight, so that, for example, the influence of the item "fast food" is significantly greater than that of the item "muesli (sweetened)". In the sample, the resulting values range from 2.26 to 6.74. With their respective values $I_{Bärbel} = 5.32$ and $I_{Klaus} = 6.05$, they are more than one standard deviation above the mean even after this calculation (M = 4.25 SD = 0.90). However, there is a clear difference in the test score of the two subjects. Bärbel's test score corresponds to a percentile rank of 88, whereas Klaus' test score even corresponds to a percentile rank of 97.

The example illustrates the differentiating effect that such weights can have. At the same time, it shows that it is precisely for this reason that weights must be chosen sensibly, as they can strongly influence results. In the present example, the dependent variable from the regression analysis would not be a good measure, as already discussed.

In summary, it can be stated that no rules of thumb can be formulated here either, but that the decisions to be made in the course of index formation must be made individually for each questionnaire. Empirical results from regression analyses can provide support and ideally even weighting factors. However, they should be used with caution and ideally supplemented by findings from the literature and from interviews with experts.

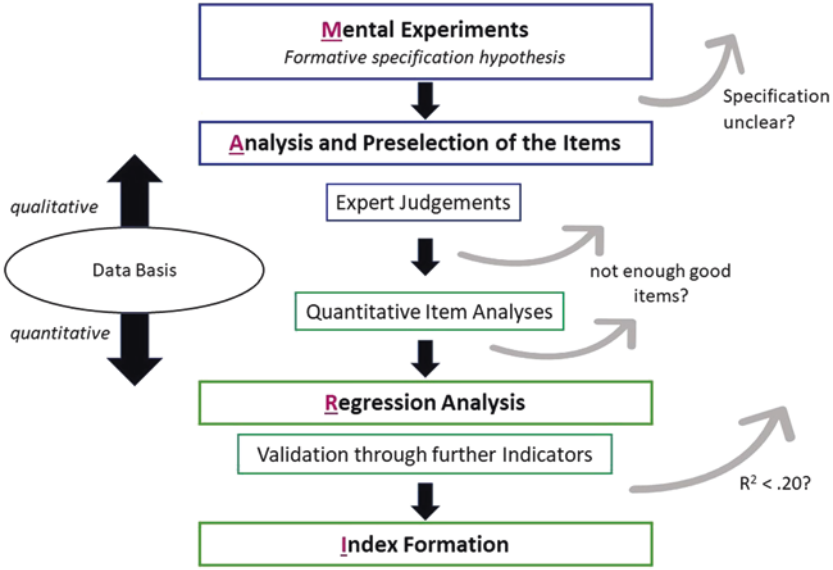


Fig. 4.3 Schematic representation of the MARI method

The following figure summarizes the procedure of the MARI method. The process of data-guided index formation must always be terminated if the results of the respective steps do not show a satisfactory fit (Fig. 4.3).

In the sense of scientific honesty and not least in order to counter the frequent criticism of the formative model, it is important to make all decisions to be made in the course of the process and the foundations on which they are based transparent and to document them in a comprehensible manner. Only then can formative modeling be carried out according to the rules of the art – lege artis.

References

Albers, S. & Hildebrandt, L. (2006). Methodische Probleme bei der Erfolgsfaktorenforschung – Messfehler, formative versus reflektive Indikatoren und die Wahl des Strukturgleichungs-Modells. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 58*(1), 2–33.

Arbuckle, J. L. (2012). *IBM SPSS Amos 21*. Chicago, IL: Amos Development Corporation.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (10., neu bearb. und erw. Aufl ed.). Springer.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016). *Multivariate Analysemethoden*. Berlin, Heidelberg: Springer.

Bainter, S. A. & Bollen, K. A. (2015). Moving Forward in the Debate on Causal Indicators: Rejoinder to Comments. *Measurement: Interdisciplinary Research and Perspectives, 13*(1), 63–74.

Binder, H., & Eberl, M. (2005). Statistisch unterstützte Spezifikationsprüfung: Die Performance von Tetrad-Test und SEM. Schriften zur Empirischen Forschung und Quantitativen Unternehmensplanung.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. (2007). Interpretational Confounding Is Due to Misspecification, Not to Type of Indicator. *Psychological Methods, 12*(2), 219–228.

Bollen, K. A., Lennox, R. D. & Dahly, D. L. (2009). Practical application of the vanishing tetrad test for causal indicator measurement models: An example from health-related quality of life. *Statistics in Medicine, 28*, 1524–1536.

Bollen, K. A. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305.

J. Riebel, H. Lichtenberg, *Formative Modelling in Psychology and Educational Science*, https://doi.org/10.1007/978-3-658-39404-2

Bollen, K. A., & Ting, K. F. (1993). Confirmatory tetrad analysis. *Sociological Methodology, 23,* 147–175.

Bollen, K. A. & Ting, K.-F. (2000). A tetrad test for causal indicators. *Psychological Methods, 5*, 3–22.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*(3), 265–284. https://doi.org/10.1037/a0024448

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic, 34*, 34–37.

Borsboom, D., Mellenbergh, G. J. & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203.

Boßow-Thies, S. & Panten, G. (2009). Analyse kausaler Wirkungszusammenhänge mit Hilfe von Partial Least Squares (PLS). In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (S. 365–380). Wiesbaden: Springer Fachmedien.

Bruhn, M., Lucco, A., & Wyss, S. (2008). Beendigung von Kundenbeziehungen aus Anbietersicht. *Marketing ZFP, 30*(4), 221–238.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research, 5*(1), 3–52.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern Methods for Business Research*, *295*(2), 295–336.

Christophersen, T., & Grape, C. (2007). Die Erfassung latenter Konstrukte mit Hilfe formativer und reflektiver Messmodelle. In S. Albers, D. Klapper, U. Konradt, A. Walter, & J. Wolf (Eds.), *Methodik der empirischen Forschung* (pp. 103–118). Gabler. https://doi.org/10.1007/978-3-8349-9121-8_8

Christophersen, T. & Grape, C. (2009). Die Erfassung latenter Konstrukte mit Hilfe formativer und reflektiver Messmodelle. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (S. 103–118). Wiesbaden: Springer Fachmedien.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Craven, R. G., Marsh, H. W., & Parada, R. H. (2013). Potent ways forward: new multidemnsional theoretical structural models of cyberbullying, cyber targetization and bystander behaviors and their potential relations to traditional bullying constructs. In S. Bauman, D. Cross, & J. L. Walker (Eds.), *Principles of cyberbullying research* (pp. 68–86). Routledge.

Diamantopoulos, A. (2006). The error term in formative measurement models: interpretation and modeling implications. *Journal of Modelling in Management, 1*(1), 7–17.

Diamantopoulos, A. & Riefler, P. (2008). Formative Indikatoren: Einige Anmerkungen zu ihrer Art, Validität und Multikollinearität. *Zeitschrift für Betriebswirtschaft, 78*(11), 1183–1196.

Diamantopoulos, A., Riefler, P. & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203–1218.

Diamantopoulos, A., & Siguaw, J. A. (2002). The Impact of Research Design Characteristics on the Evaluation and Use of Export Marketing Research: An Empirical Study. *Journal of Marketing Management, 18*(1–2), 73–104. https://doi.org/10.1362/0267257022775936

Diamantopoulos, A. & Winklhofer, H. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277.
Dlugosch, G.E. & Krieger, W. (1995). *Fragebogen zur Erfassung des Gesundheitsverhaltens*. Frankfurt: Swets & Zeitlinger.
Eberl, M. (2004). *Formative und reflektive Indikatoren im Forschungsprozess: Entscheidungsregeln und die Dominanz des reflektiven Modells*. Ludwig-Maximilians-Universität München: Schriften zur Empirischen Forschung und Quantitativen Unternehmensplanung.
Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*(2), 370–388.
Edwards, J. R. & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–174.
Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden: Lehrbuch. Mit Online-Material* (4., Originalausgabe, 4., überarbeitete und erweiterte Aufl ed.). Weinheim: Beltz.
Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (Eds.). (2010). *Handbook of Partial Least Squares*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-32827-8.
Festl, R. (2015). *Täter im Internet: Eine Analyse individueller und struktureller. Erklärungsfaktoren von Cybermobbing im Schulkontext*. Wiesbaden: VS Verlag für Sozialwissenschaften.
Fluck, J. (2020a). *Cyberbullying – Theoretische und empirische Analysen zur Konstruktklärung und Messmodellierung eines Gewaltphänomens.* Hamburg: Dr. Kovac.
Fluck, J. (2020b). Formative Messmodelle – Eine nützliche Ergänzung für das Methodenrepertoire zur Testkonstruktion in Psychologie und Erziehungswissenschaft. *Empirische Pädagogik*, 34(2), 179–188.
Fornell, C. & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19*, 440.
Fornell, C. & Cha, J. (1994). Partial least squares. In R. P. Bagozzi (Ed.) *Advanced Methods of Marketing Research* (pp. 52–78). Cambridge: Blackwell.
Fuchs, A. (2011). *Methodische Aspekte linearer Strukturgleichungsmodelle. Ein Vergleich von kovarianz- und varianzbasierten Kausalanalyseverfahren* (Vol. 2/2011): Julius-Maximilians-Universität Würzburg, Lehrstuhl für BWL und Marketing, Würzburg.
Funke, E. M., Lorenzi, S. & Winkler, R. S. (2020). Entwicklung eines Fragebogens zur Erfassung der Zufriedenheit von Studierenden mit dem Masterstudiengang Empirische Bildungsforschung. Projektbericht im Seminar Forschungswerkstatt Quantitative Verfahren. RWTH Aachen: Institut für Erziehungswissenschaft.
Geiser, C. (2011). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung*. VS Verlag für Sozialwissenschaften.
Gudergan, S. P., Ringle, C. M., Wende, S. & Will, A. (2008). Confirmatory tetrad analysis in PLS path modeling. *Journal of Business Research*, *61*(12), 1238–1249.
Häder, M. (2009). *Delphi-Befragungen*. (2 ed.). VS Verlag für Sozialwissenschaften. https://doi.org/https://doi.org/10.1007/978-3-531-91926-3
Hamby, S. L., & Finkelhor, D. (2000). The Victimization of Children: Recommendations for Assessment and Instrument Development. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(7), 829–840. https://doi.org/https://doi.org/10.1097/00004583-200007000-00011

Hardin, A. M., Chang, J. C.-J., Fuller, M. A. & Torkzadeh, G. (2011). Formative Measurement and Academic Research: In Search of Measurement Theory. *Educational and Psychological Measurement, 71*(2), 281–305.

Hauser, R. M. (1971). *Socioeconomic background and educational performance*. Washington: American Sociological Association.

Hauser, R. M., & Goldberger, A. S. (1971). The Treatment of Unobservable Variables in Path Analysis. *Sociological Methodology, 3*, 81–117. https://doi.org/10.2307/270819

Henseler, J., & Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics, 28*(2), 565–580. https://doi.org/10.1007/s00180-012-0317-1

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A. & Straub, D. W. (2014). Common beliefs and reality about PLS: Comments on Ronkko and Evermann (2013). *Organizational Research Methods, 17*, 182–209.

Hipp, J. R. (2008). *Performing Vanishing Tetrads Tests Using CTANEST1*. Verfügbar unter: https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCIQFjAAahUKEwiK0afTx4_IAhWmh3IKHZTKBys&url=https%3A%2F%2Fwebfiles.uci.edu%2Fhippj%2Fjohnhipp%2FCTANEST1_documentation.doc&usg=AFQjCNGAWdHQs-2y4_2Rs_kCPh-oZ2lR6g&sig2=FtmYFo79tGlQLiMuiaGqQQ&bvm=bv.103627116,d.bGQ&cad=rja [28.10.2018].

Hipp, J. R., Bauer, D. J. & Bollen, K. A. (2005). Conducting tetrad tests of model fit and contrasts of tetrad-nested models: A new SAS macro. *Structural Equation Modeling: A Multidisciplinary Journal, 12*, 76–93.

Homburg, C., Hoyer, W. D., & Fassnacht, M. (2002). Service Orientation of a Retailer's Business Strategy: Dimensions, Antecedents, and Performance Outcomes. *Journal of Marketing, 66*(4), 86–101. https://doi.org/10.1509/jmkg.66.4.86.18511

Howell, R. D., Breivik, E. & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12*(2), 205.

Hu, L.-T. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification, *Psychological Methods, 3*, 424–453.

Ihme, J. M., & Senkbeil, M. (2017). Warum können Jugendliche ihre eigenen computerbezogenen Kompetenzen nicht realistisch einschätzen? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 49*(1), 24–37. https://doi.org/10.1026/0049-8637/a000164

Jarvis, C. B., MacKenzie, S. B. & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199–218.

Jäger, R. S. & Petermann, F. (1995). *Psychologische Diagnostik: Ein Lehrbuch* (3., korr. Aufl). Weinheim: Beltz, Psychologie Verlags Union.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American Statistical Association, 70*(351a), 631–639. https://doi.org/10.1080/01621459.1975.10482485

Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Lincolnwood, IL: Scientific Software International.

Kim, G., Shin, B. & Grover, V. (2010). Investigating two contradictory views of formative measurement in information systems research *MIS Quarterly, 34*(2), 345–A345.

Klauer, K. J. (1987). *Kriteriumsorientierte Tests.* Göttingen: Hogrefe.
Konradt, U., Christophersen, T., & Ellwart, T. (2008). Erfolgsfaktoren des Lerntransfers unter computergestütztem Lernen. *Zeitschrift für Personalpsychologie, 7*(2), 90–103. https://doi.org/10.1026/1617-6391.7.2.90
Lohmüller, J. B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg: Physica-Verlag.
Markus, K. A. (2014). Unfinished business in clarifying causal measurement: Commentary on Bainter and Bollen. *Measurement: Interdisciplinary Research and Perspectives, 12*, 146–150.
Moosbrugger, H. & Kelava, T. (2012). *Testtheorie und Fragebogenkonstruktion.* Berlin: Springer.
Moosmüller, G. (2004). *Methoden der empirischen Wirtschaftsforschung*. München: Pearson Studium.
Muthén, L. K. & Muthén, B. O. (2010). *1998–2010 Mplus user's guide*. 6th Edition. Los Angeles, CA: Muthén & Muthén.
Riebel, J. (2010). Modellierungskompetenzen beim mathematischen Problemlösen. Universität Koblenz-Landau: Dissertation. Verfügbar unter: https://kola.opus.hbz-nrw.de/opus45-kola/frontdoor/deliver/index/docId/397/file/Publikation.pdf [20.10.2020].
Ringle, C. M., Wende, S. & Becker, J.-M. (2015). *SmartPLS 3. Boenningstedt: SmartPLS GmbH.* Verfügbar unter: https://www.smartpls.com [28.10.2018].
Ringle, C. M., Wende, S. & Becker, J.-M. (n.d.). *The standardized root mean square residual (SRMR) is a goodness of (model) fit measure for PLS-SEM*. Verfügbar unter: https://www.smartpls.de/documentation/srmr [20.10.2020].
Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International journal of research in marketing, 19*(4), 305–335.
Schmidt-Atzert, L., & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl.). Berlin: Springer
Schmitz, L., Brodesser, E., & Pant, H. A. (2020). Adaptive Lehrkompetenz: Bildung von Indizes und empirische Ergebnisse zur Wirkung universitärer Lehrveranstaltungen. In *Inklusionsorientierte Lehr-Lern-Bausteine für die Hochschullehre. Ein Konzept zur Professionalisierung zukünftiger Lehrkräfte* (pp. 124–136). Verlag Julius Klinkhardt. https://doi.org/10.25656/01:19023, https://doi.org/10.35468/5798_04.2
Schnell, R., Hill, P. B. & Esser, E. (2018). *Methoden der empirischen Sozialforschung.* Berlin: De Gruyter Oldenbourg
Schneider, H. (2009). Nachweis und Behandlung von Multikollinearität. In *Methodik der empirischen Forschung* (pp. 221–236). Springer.
Simonetto, A. (2012). Formative and reflective models: State of the art. *Electronic Journal of Applied Statistical Analysis, 5*(3), 452–457.
Smith, T.W. (2011). Measurement in Health Psychology Research. In H.S. Friedman (Ed.). *The Oxford Handbook of Health Psychology* (pp. 42–72). New York: Oxford University Press.
Steffen, A. (1994). *Das Problem der Multikollinearität in Regressionsanalysen*. Frankfurt a. M.: Peter Lang.
Steyer, R. & Eid, M. (2001). *Messen und Testen. Mit Übungen und Lösungen.*Berlin: Springer.

Taylor, S. E. (2011). Social support: A review. In H. S. Friedman (Ed.), *The Oxford handbook of health psychology* (pp. 189–214). New York: Oxford University Press.

Temme, D. (2006). Die Spezifikation und Identifikation formativer Messmodelle der Marketingforschung in Kovarianzstrukturanalysen. *Marketing ZFP, 28*(3), 183–196.

Tepper, B. J., & Henle, C. A. (2011). A case for recognizing distinctions among constructs that capture interpersonal mistreatment in work organizations. *Journal of Organizational Behavior, 32*(3), 487–498.

Tu, W., & Liu, H. (2002). Zero-inflated data. *Wiley StatsRef: Statistics Reference Online*.

Weiber, R., & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung: Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS* (2., erw. und korrigierte Aufl ed.). Springer Gabler.

West, S. G. & Grimm, K. J. (2014). Causal Indicator Models: Unresolved Issues of Construction and Evaluation. *Measurement: Interdisciplinary Research and Perspectives, 12*(4), 160–164.