MEASUREMENT IN MARKETING

Hans Baumgartner and Bert Weijters

Suggested Citation: Hans Baumgartner and Bert Weijters (2019), "Measurement in Marketing", Foundations and Trends® in Marketing: Vol. 12, No. 4, pp 278–400. DOI: 10.1561/1700000058.

Hans Baumgartner is Smeal Professor of Marketing in the Smeal College of Business at The Pennsylvania State University, 482 Business Building, University Park, PA 16802, Phone: (814) 863-3559, E-mail: hansbaumgartner@psu.edu. Bert Weijters is Associate Professor in the Department of Personnel Management and Work and Organizational Psychology, Ghent University, Dunantlaan 2, B-9000 Ghent, Belgium, E-mail: bert.weijters@ugent.be. The authors thank Zac Rolnik, the editors, and two reviewers for helpful comments.

Contents

ABSTRACT

We distinguish three senses of the concept of measurement (measurement as the selection of observable indicators of theoretical concepts, measurement as the collection of data from respondents, and measurement as the formulation of measurement models linking observable indicators to latent factors representing the theoretical concepts), and we review important issues related to measurement in each of these senses. With regard to measurement in the first sense, we distinguish the steps of construct definition and item generation, and we review scale development efforts reported in three major marketing journals since 2000 to illustrate these steps and derive practical guidelines. With regard to measurement in the second sense, we look at the survey process from the respondent's perspective and discuss the goals that may guide participants' behavior during a survey, the cognitive resources that respondents devote to answering survey questions, and the problems that may occur at the various steps of the survey process. Finally, with regard to measurement in the third sense, we discuss both reflective and formative measurement models, and we explain how researchers can assess the quality of measurement in both types of measurement models and how they can ascertain the comparability of measurements across different populations of respondents or conditions of measurement. We also provide a detailed empirical example of measurement analysis for reflective measurement models.

1

Introduction

Measurement is indispensable for empirical research in marketing, and researchers who have conducted empirical studies will have at least a rudimentary understanding of what measurement entails. Still, the concept of measurement is difficult to define unambiguously, and existing definitions (e.g., Stevens, 1946), although often cited, have been criticized on various grounds. Instead of offering yet another definition, which would probably be subject to criticism as soon as it was proposed, we will distinguish three related but distinct senses in which one can think about measurement. Based on this classification, we will then discuss issues relevant to each notion of measurement.

In one sense measurement means conceptualizing theoretical variables of interest and choosing appropriate observable indicators of the intended construct. In another sense measurement means collecting the data necessary for an empirical examination of the theoretical issues under study. In a final sense measurement means constructing a model that relates the data collected in the second step to the latent factors representing the concepts the researcher is interested in, as specified in the first step. Sometimes, it is difficult to clearly distinguish the three activities, as when a researcher employs existing data to study an issue and uses single observed variables as approximations of presumed theoretical concepts of interest. At other times, multiple observed indicators of carefully defined constructs are developed, primary data from specially chosen respondents are carefully collected, and sophisticated measurement models are formulated to maximize the correspondence between the observed responses and the latent concepts of interest.

The primary goal of this monograph is to review important issues related to measurement in all three senses. To supplement the theoretical discussion, we will present empirical data on how recent research published in three important marketing journals (*Journal of Consumer Research [JCR]*, *Journal of Marketing [JM]*, and *Journal of Marketing Research [JMR]*) has dealt with some of these issues (with an emphasis on measurement in the first sense), and we will also report a detailed example of measurement analysis in the context of material values.

Measurement is intimately related to construct validity and procedures for assessing the construct validity of measures. Construct validity is commonly viewed as the extent to which the measures designed to operationalize abstract theoretical concepts approximate the constructs in question (Bagozzi, 1980; Churchill, 1979; MacKenzie, Podsakoff, and Podsakoff, 2011; Peter, 1981). A prerequisite for establishing construct validity is that theoretical concepts be defined clearly and that empirical operationalizations accurately capture all the facets, and only the relevant facets, of the intended construct. These issues relate most closely to the first sense of measurement and are discussed in section 2. Assessing the construct validity of measures also entails procedures for ascertaining the reliability and convergent and discriminant validity of measures of the construct(s) of interest (Campbell and Fiske, 1959), including efforts to demonstrate that observed measures are not seriously contaminated by sources of systematic variance unrelated to the intended construct (particularly so-called method effects; Podsakoff *et al*., 2003). These issues are discussed extensively in section 4 in the context of measurement in the third sense. Since constructs are theoretical entities (regardless of whether they are assumed to be figments of the researcher's imagination or thought to exist in the real world), many authors have suggested that an important part of construct validation is that observed measures behave as expected by a theory in which the construct of interest plays a prominent role

(Bagozzi, 1980, 1984; Churchill, 1979; Nunnally 1978). In other words, a measure should have nomological validity by fitting into a nomological net of related constructs as specified by some theory. Although we agree that nomological validity is an important aspect of construct validity, we will not emphasize this aspect because assessing the nomological validity of a measure is dependent on a particular theory and thus difficult to discuss in the abstract. Furthermore, nomological validity tests are beyond the scope of measurement analysis per se.

Before we proceed, several comments are in order. First, a discussion of measurement could easily fill a tome, and we had to make decisions, based on our own preferences, about what should be included in this monograph. We hope readers will agree with our selections and find the discussion helpful. Second, although measurement need not necessarily involve the assignment of numbers to objects and events, we will focus on this type of measurement. Third and closely related to the previous point, the treatment of measurement is restricted to what has been called the psychometric approach to measurement (usually based on rating scales), in contrast to the representational approach (Judd and McClelland, 1998). The reason is that we believe this approach is most useful to the practicing empirical researcher. Fourth, there are different modes of data collection (observation, interviews, questionnaires, etc.), and there are unique issues that arise when using each of these data collection methods. Our focus will be on survey data collection methods (in a broad sense) via questionnaires (including internet surveys) because these are most common in marketing. Fifth, when we mention examples of prior measurement practices and offer critical reflections, our intention is not to disparage previous work, but to offer tangible illustrations of the points we are trying to make, with the hope of improving future research practices.

2

Measurement as the selection of observable indicators of theoretical concepts

The canonical approach to measurement in the first sense consists of (1) defining the theoretical construct(s) of interest and (2) selecting observable measures of the theoretical construct(s). We call this the *construct-first approach*. Not all measurement proceeds in this way. Sometimes researchers use what one might call the *measure-first approach*. This is particularly common in secondary research in which the data were not collected for the purpose that the researcher has in mind, and existing data sources are used serendipitously to study a phenomenon of interest. We will briefly discuss the latter approach and then turn to the former, which is generally the preferred way of measuring phenomena.

2.1 Measure-first approach

With this approach, the selection of observable measures is not primarily guided by theoretical considerations. Rather, the researcher tries to make the best of the measures that happen to be available. As a consequence, observed measures are usually at best proxy variables for what the researcher has in mind, and frequently multiple measures are unavailable so that it is difficult to assess basic desiderata of measurement quality such as reliability and convergent validity. The researcher's task becomes one of justifying the use of existing data as measures of presumed theoretical concepts and to present at least suggestive evidence about hypothesized conceptual relationships between constructs based on empirical associations among fallible observed measures.

A recent study by Viswanathan, Li, John, and Narasimhan (2018) serves as an illustration of this approach. In a field study, salespeople's incentive compensation was changed from a

combination of merchandise and cash incentives to a pure cash incentive scheme. The pure cash program reduced sales, and the authors wanted to show that reduced effort mediated the effect of the change in the incentive plan on sales. However, no self-report data were available for the construct of salesperson effort, in part because the authors thought that measuring effort might introduce demand artifacts. The change in effort was therefore inferred from the salesperson fixed effects in the sales equation while accounting for various other variables such as seasonality. It is impossible to conduct a detailed measurement analysis in this situation, and there is no guarantee that the presumed effort variable actually measures salesperson effort.

2.2 Construct-first approach

With this approach, the measurement process starts with a clear conception of what is to be measured, and the measures are purposely chosen to capture the theoretical entity of interest as accurately as possible. Since developing good measures of intended constructs is difficult, it is ideal if validated measurement instruments are readily available. Scale handbooks such as Bearden, Netemeyer, and Hawes (2011) or Bruner II (2015) can be useful for selecting measures that have been developed in prior research. Unfortunately, researchers sometimes rely on scales that are only somewhat similar in content to the desired construct, and too often the primary justification for using a certain scale is that some other researcher has previously used it.

The prototype of the construct-first approach to measurement is the scale development process described in sources such as Churchill (1979), DeVellis (2003), MacKenzie, Podsakoff, and Podsakoff (2011), and Netemeyer, Bearden and Sharma (2003), among others. The process consists of two basic steps (Baumgartner and Weijters, 2017). In the first step, the researcher carefully defines the construct, or conceptual entity, that he or she is interested in. This involves

explicating the essential meaning of the construct and distinguishing it from other, related constructs. In the most general sense, the researcher must specify the domain of the construct in terms of (a) the attribute(s) to be measured, or the properties and characteristics that are the constitutive elements of the concept the researcher has in mind, (b) the object of measurement, or the entity or entities to which the attributes are ascribed, and (c) the rater, or the provider of the judgment (see also Rossiter, 2002). For example, for the construct of need for touch (Peck and Childers, 2003), the attribute to be measured is preference for haptic information, the object of measurement is the consumer, and the rater is the consumer as well (assuming that need for touch is based on self-report). As another example, for the construct of perceived service quality (Brady and Cronin, 2001), the attribute to be measured is an evaluation of service performance, the object of measurement is the service provider, and the rater is the customer. The attributes to be measured are generally restricted to a certain context. For example, a consumer's need for uniqueness refers to differentiation through the acquisition, utilization, and disposition of consumer goods, not need for uniqueness in other life domains (see Tian, Bearden, and Hunter, 2001). Also, attributes can refer to properties of a relationship between the object of measurement and a particular target. For example, in the case of brand love, the attribute is a consumer's love relationship with a specific brand. However, the object of measurement is the consumer, not the brand, and the attribute is the consumer's deep commitment to a particular brand. Usually, there are constructs that are similar, or at least somewhat related, to the concept of interest, so to get a clearer understanding of the defining features of the construct, it is useful to compare and contrast the intended construct with constructs that exhibit a family resemblance.

Coming up with an appealing conceptualization of a construct is partly an art, not just a science, although reviews of prior writings on the topic or qualitative research with domain

experts are useful in guiding the conceptual delineation of the concept of interest. In a best-case scenario, a researcher may be able to draw upon a theory that provides explicit guidance on how to think about the essential elements of a construct. For example, if the researcher is interested in the construct of perceived spokesperson credibility in a persuasion setting, the work of Hovland, Janis, and Kelley (1953) may suggest that source credibility involves both the extent to which a source is believed to be capable of making correct assertions (expertise) and the extent to which the source is believed to consider his or her assertions to be valid (trustworthiness). Since credibility is a characteristic of the source of a persuasive communication, it also has to be distinguished from other source variables such as source attractiveness and source power (see Pornpitakpan, 2004).

One important issue that must be considered when defining a construct is whether the construct should be conceptualized as unidimensional or multi-dimensional. In principle, the issue of dimensionality refers to both the object of measurement and the attributes of the object to be measured. For example, if the construct is a service provider's service quality, the object at the highest level of abstraction may be the service provider, but at a lower level the more specific object could be the firm's employees, the physical environment, or the product/service offered by the firm (e.g., Brady and Cronin, 2001). The attributes could be different aspects of service quality, such as reliability, responsiveness, or empathy (as in the SERVQUAL conceptualization; see Parasuraman, Zeithaml, and Berry, 1988). As a general principle, one can say that a multi-dimensional conceptualization becomes more likely (and probably more necessary) when constructs are more abstract than concrete (either in terms of the attributes that describe the construct or the range of objects to which the attributes extend; see Rossiter, 2002). Although the issue of dimensionality should be considered for both the object and attributes of the construct, it

is usually most relevant for the conceptualization of the construct's attributes since the object of the construct is often relatively concrete.

When a construct is treated as multi-dimensional, the question arises whether different hierarchical levels of the construct should be distinguished. Often, researchers simply assume multiple correlated factors representing the dimensions of the construct without specifying a higher-order factor structure. However, the conceptualization is more complete when a more specific super-structure is imposed. If a higher-order structure is hypothesized, a researcher must specify how many higher-order factors and how many levels there are, and whether the higher-level factors cause the lower-level factors or whether the higher-level factors are caused by the lower-level factors. In other words, the researcher must decide whether the lower-level factors are reflections of the higher-level factors or whether the lower-level factors form the higher-level factors (e.g., MacKenzie *et al*., 2011). Figure 2.1 depicts some of the possible factor structures that a researcher may consider, assuming there are four first-order factors and one or two second-order factors.

In the second step, the researcher must develop operational measures of the construct (or dimensions of the construct) of interest. The items generated should fully capture the intended conceptual entity, but not overlap (too much) with related but presumably distinct constructs. Item generation is usually based on a combination of methods such as inspection of prior scales measuring the same or related constructs, reviews of academic and popular literature on the topic, brainstorming and introspection, and qualitative research with respondents who are knowledgeable about the construct. Researchers usually generate many potential items and then prune the original item pool using a variety of procedures. For example, experts might be consulted to judge the content validity of each item, based on a definition of the construct or its

individual dimensions, or respondents are asked to allocate each item to one of the dimensions or rate the items in terms of their applicability or relevance to each dimension. Items that are not consistently classified as measuring the intended dimension are eliminated. In addition, items are also screened for clarity in wording and formulation. More sophisticated scale development techniques, usually based on exploratory or confirmatory factor analysis, are used as well, once a manageable item pool has been arrived at, but these techniques will be discussed below under the third sense of measurement, because they involve formulating measurement models and assessing the correspondence between a construct (more specifically, a factor thought to represent a construct) and its measures, as well as the relationship of the target construct with related constructs.

Table 2.1 reports illustrative examples and descriptions of new scale development efforts reported in JCR, JM, and JMR since 2000. The table contains the following information about each scale development project: the proposed construct; the definition of the construct (if provided in the paper); the dimensions of the construct; sample items for each (major) dimension of the construct; the measurement model used to link the observed measures to the underlying construct (to be discussed in more detail in section 4); and a characterization of the scale in terms of the object and attribute(s) being quantified, as well as the source of the quantification (a rater in all cases).

Most of the scales (18 out of 25) assess individual differences in some attribute (i.e., the object of measurement is the individual, either a consumer or a salesperson), but other objects of measurement are the firm or SBU (4) and the product or brand (3). Six scales were conceptualized as being uni-dimensional, or ultimately resulted in a one-dimensional scale, but most scales were developed to assess a multi-dimensional construct, with the number of (first-

order) dimensions ranging from 2 to 14 in the final scale (although in one paper 23 factors were considered at first). In 11 instances a higher-order factor structure was specified, or directed relationships were specified between the dimensions. Most higher-order factor structures were assumed to be reflective, but in two instances a formative model was hypothesized. The number of items in the scale ranged from 4 to 78. Only four scales contained reversed items; in one instance the direction of the response scale may have been reversed for one item.

Based on our review of these scale development efforts, we offer the following recommendations. First, the definition of a construct should include both a specification of the object to which the construct refers and the attributes that characterize the construct in question. Furthermore, there should be consistency in how these aspects of the construct are defined and measured. This was not always the case. For example, if a researcher is interested in the hedonic and utilitarian dimensions of consumers' attitudes toward products and brands, the construct presumably refers to two types of consumer reactions to products and brands (i.e., two attributes of the consumer as an object), so the scale should *not* measure whether the product or brand is, say, effective or exciting (i.e., attributes of the product or brand as an object; see Voss *et al.*, 2003). Furthermore, it is problematic if the object of the construct differs across items. For example, when measuring brand love, the object is the consumer and the attributes are various aspects of consumers' relationships with brands. However, in the brand love scale, some items refer to consumers' love of brands (e.g., I feel a sense of long-term commitment to this brand, I feel myself craving to use the brand), whereas others refer to the brand (e.g., the brand makes you look like what you want to look; Batra *et al.*, 2012). Similarly, when measuring consumer preference for local food or locavorism, the object is the consumer and the attribute is the

preference, but some items in the locavorism scale have foods, food systems or producers as their object and focus on attributes such as taste, quality or societal impact (Reich *et al*., 2018).

Second, researchers sometimes focus too much on completeness (i.e., measuring all relevant content within the domain of interest being covered by the scale) at the expense of parsimony (i.e., avoiding redundancy in the number of items and factors). Although a unidimensional scale is not always achievable, particularly for abstract and multi-faceted constructs, some measurement instruments are overly complex. For example, the conceptual appeal of a 14-factor scale in which brand love is specified as a third-order factor model indicated by a mix of first- and second-order factors is debatable (Batra *et al*., 2012). A scale consisting of 57 items is probably too long to include in most questionnaires, and complex factor structures like this are unlikely to be upheld in future studies.

Third, in direct contrast to the previous point and probably encouraged (at least in part) by the use of confirmatory factor analysis for validating scales (which imposes very stringent and empirically unrealistic requirements on the factor space, such as zero loadings on non-target factors), researchers increasingly propose rather simplistic measurement scales in which the different items measuring a given dimension are minor linguistic variations of the same statement. As just one example, the aesthetics dimension of the new product design construct (Homburg, Schwemmle, and Kuehnl, 2015) is measured by the following three items: The product is visually striking; the product is good looking; and the product looks appealing. While such a brief scale is parsimonious and convenient to use, one wonders whether the aesthetics of product design can be captured adequately with such items. In other words, it seems questionable whether the indicators fully cover the domain of interest.

Fourth, although the distinction between reflective and formative indicators has been discussed at length in the literature (MacKenzie, Podsakoff, and Jarvis, 2005), in many cases it is not straightforward to decide how the relationship between different hierarchical dimensions of a construct should be specified. For example, on the one hand one could argue that dimensions such as product standardization, promotion standardization, or standardized channel structure are contributing factors to the construct of Global Marketing Strategy (Zou and Cavusgil, 2002). On the other hand, the decision to standardize these marketing mix components is probably taken deliberately and is thus reflective of the extent to which management aims for a globally standardized approach. In other instances, it may be relatively clear that a measurement model should be reflective. For example, Reich *et al*. (2018) logically explain why locavorism (i.e., consumer preference for local food) and its dimensions are measured with reflective indicators by explicitly evaluating the criteria proposed by Jarvis *et al*. (2003).

In summary, the first step in the measurement process consists of conceptualizing the theoretical entity or entities that the researcher is interested in and developing items (usually statements or questions to which raters are asked to respond) that capture the essence of the (usually) intangible phenomena in researchers' explanatory frameworks. The goal is to come up with a parsimonious description of the constructs of interest that clarifies their meaning (in terms of what a construct represents and how it differs from related constructs) and forms the basis for developing observed measures that enable empirical investigations.

3

Measurement as the collection of data from respondents

Collecting data from respondents includes many important decisions, such as defining the target population and choosing a sampling frame, drawing a sample, and taking steps to minimize nonresponse errors (what Groves *et al.*, 2004, call survey unit representation). Although these issues are important, they do not directly deal with measurement per se. Here, we will focus instead on the task that respondents have to complete when they are asked to respond to a survey.

There are three important components of a model of how respondents answer questions in surveys, which determine the quality of the resulting data. First, what are the goals that guide participants' behavior during a survey (e.g., do respondents primarily want to provide accurate information or make a good impression)? Second, are respondents willing and able to devote sufficient cognitive resources to a survey to provide accurate responses (or are they, for instance, not paying attention)? Third, which tasks do respondents have to execute to provide answers to survey questions (in terms of comprehension, retrieval, etc.), and what are some common threats to survey validity at each step?

Researchers usually assume that respondents have an accuracy goal when completing a survey and are both able and motivated to provide accurate responses. This is not necessarily the case, and we will consider both the situation in which respondents hold a goal other than accuracy and the situation in which respondents lack the motivation and/or ability to respond accurately (Baumgartner and Weijters, 2012). In addition, we will briefly discuss important issues that arise at the various steps of the cognitive process that respondents go through when answering questions.

3.1 Respondents' goals: Accuracy vs. self-presentation

Researchers generally assume that when respondents complete a survey, they are guided by the goal of accuracy. Accuracy refers to a desire to provide responses that are a faithful representation of reality. For matters of fact, the response rendered will ideally be objectively accurate, but even when this is not entirely the case, the respondent hopefully tries, to the best of his or her ability, to be accurate. For matters of opinion, the response should accurately reflect a participant's true belief. Accuracy requires that the comprehension of the question, the formation of an (internal) response, and the communication of an overt answer be evenhanded, impartial, and open-minded.

Unfortunately, survey participants are sometimes guided by goals other than accuracy, which are likely to contaminate the conclusions derived from a survey (Baumgartner and Weijters, 2012). Among these extraneous goals, the most important is probably self-presentation. In a survey setting, self-presentation refers to respondents' desire to provide answers that make them look good to others. This is usually referred to as socially desirable responding (Steenkamp, De Jong, and Baumgartner, 2010). Targets of self-presentation include the interviewer, the sponsor of the survey, people who are present during the administration of the survey, or the public at large. In contrast to accuracy, the comprehension of the question, the formation of an (internal) response, and the communication of an overt answer is biased, partial, and possibly strategic.

While the notion of socially desirable responding (SDR) is old, our understanding of this concept has changed considerably in recent years (see Steenkamp, De Jong, and Baumgartner, 2010, for a recent review). Traditionally, SDR was conceptualized and measured as a

unidimensional construct. However, the most recent conceptualization distinguishes four varieties of SDR based on a cross-classification of two dimensions: domain of content (agency vs. communion) and degree of awareness (conscious vs. non- or sub-conscious; see Paulhus, 2002). With respect to the former, respondents may (attempt to) present themselves as either superheroes in the agency domain (reflecting a desire to be autonomous, dominant, and unique) or as saints in the communion domain (reflecting a desire to feel connected, belong, and seek approval). With respect to the latter, respondents may pursue these goals deliberately or (more or less) subconsciously.

Although certain topics may be conducive to SDR across people, research has shown that there are reliable individual differences in SDR, and several instruments have been developed to measure people's general tendency to engage in SDR. These scales are based on the idea that those high in SDR will exaggerate uncommon desirable behaviors and deny common undesirable behaviors. The most commonly used scale is the Marlowe-Crowne social desirability scale (Crowne and Marlowe, 1964), but it is a unidimensional scale that confounds superhero and saint-like inclinations (Steenkamp *et al*., 2010). The most sophisticated scale is the balanced inventory of desirable responding (BIDR; Paulhus, 1991), which consists of two dimensions, SDE (self-deceptive enhancement) and IM (impression management). SDE and IM were originally thought to assess the nonconscious and conscious forms of positivity bias, respectively, but there is now evidence that they actually measure superhero and saint-like tendencies. Paulhus (2002) goes further and proposes that SDE measures unconscious superhero bias, whereas IM measures deliberate saint-like inclinations, but there is little empirical support for this suggestion (Steenkamp *et al*., 2010).

The fact that individual differences in SDR can be measured suggests that one can use an SDR scale to control for possible contamination of observed responses to measures of substantive constructs (e.g., by using respondents' scores on an SDR scale as a covariate). However, some caution is required when using this procedure (Steenkamp *et al*., 2010). Research has shown that the agency and community dimensions of SDR are nomologically related to a constellation of personality traits and values, as well as dimensions of national culture, which implies that a correlation of a substantive scale with SDR does not necessarily imply contamination, but instead may reflect a substantive relationship. For example, if the construct of interest is social approval, a relationship with IM (as a measure of the communion dimension of SDR) is to be expected.

Instead of thoughtlessly correlating substantive scales with a measure of SDR (presumably because "everybody knows that such a correlation is bad"), researchers must think carefully about whether SDR is likely to contaminate observed measures. We suggest that researchers ask themselves the following questions (Steenkamp *et al*., 2010). First, is the measurement context a so-called high-demand situation? Research has shown that when surveys ask questions about sensitive topics, when there is a possibility of public disclosure of responses (esp. to sensitive questions), and when important outcomes are at stake, extra caution is required because high-demand situations encourage SDR (Tourangeau and Yan, 2007). In particular, high-demand situations are more likely to lead to deliberate response management, which is usually of greater concern than unconscious SDR. In fact, one may question whether unconscious SDR is problematic at all since the exaggerated self-descriptions rendered in this case are sincere. Second, does the respondent want to project a favorable self-image in the agency or communion domain in the high-demand situation under consideration? Depending on

the answer to this question, the SDE or IM scale should be used to assess individual differences in SDR. Finally, is the correlation between the substantive scale(s) and the relevant SDR scale large enough to cause concern? It is difficult to give general guidelines on what counts as a worrisome level of correlation, but Steenkamp *et al*. (2010) suggest that a standardized regression coefficient exceeding .2 in absolute magnitude might signal potential problems. Steenkamp *et al*. also present illustrative data about the correspondence between SDE and IM and 9 different substantive scales (e.g., ethnocentrism, health consciousness, deal proneness) in 26 different countries in Europe, Asia, and North and South America. Their data suggest that different constructs are differentially related to superhero and saint-like tendencies and that there is substantial variation in the size of correlations across different countries.

3.2 Respondents' ability and motivation to answer questions accurately: Optimizing vs. satisficing

While goals determine the direction of survey respondents' engagement in the survey task, their ability and motivation influence the intensity of goal pursuit. In particular, even if respondents have an accuracy goal, the quality of their responses might be poor if they are not sufficiently able and motivated to provide accurate responses. Both dispositional and situational factors can lead to variation in ability and motivation (Petty *et al*., 2005). Examples of dispositional factors that affect a person's ability and motivation to provide accurate responses include verbal ability and need for cognition, respectively. Examples of situational factors that affect a person's ability and motivation to provide accurate responses include distractions in the survey setting and the length of the survey, respectively.

When respondents are able and willing to devote sufficient effort to a task in order to provide an accurate response, they are said to be optimizing. However, consistent with the view of people as cognitive misers (Fiske and Taylor, 1991), respondents often minimize the amount of cognitive resources invested in formulating a response to questionnaire items, and they may choose a response that is 'good enough' rather than optimal. If this is the case, they are said to be satisficing (Krosnick, 1991; see also MacKenzie and Podsakoff, 2012). Satisficing is conceptually similar to Meade and Craig's (2012) notion of carelessness or inattentiveness and Huang *et al*.'s (2012) concept of insufficient effort responding. Satisficing and optimizing should not be thought of as a dichotomy but as the endpoints of a continuum reflecting differences in the effort that respondents are able and willing to devote to completing a survey and providing accurate responses.

Since satisficing is detrimental to survey quality, researchers usually try to encourage respondents to optimize. This can be done by emphasizing that the research is important, asking respondents to take the task seriously and answer truthfully, and possibly providing incentives for participation (in hopes that these incentives will be reciprocated with good behavior). As an alternative to these *a priori* procedural strategies, researchers often try to assess satisficing behavior after the fact. If satisficing can be measured successfully *a posteriori*, respondents who do not put sufficient effort into the response task can be eliminated, or measures of satisficing can be used to control for careless responding statistically.

Different methods for identifying satisficing (and those who engage in it) have been discussed in the literature (Huang *et al*., 2012; Meade and Craig, 2012). We will differentiate these approaches based on two dimensions: (a) whether special measures are included in the survey that are specifically designed to assess satisficing, or whether satisficing is inferred from

the way respondents answer the substantive items in the survey; and (b) whether satisficing and its underlying causes are assessed directly, or whether the measurement of satisficing is based on the presumed consequences of satisficing (see Table 3.1).

One method of assessing satisficing directly via specially constructed scales is to ask respondents how much effort they spent on answering the questions in a survey (category 1 in Table 3.1). For example, Meade and Craig (2012) developed a 15-item scale of Participant Engagement, which consists of two factors, Diligence (e.g., I carefully read every survey item, I worked to the best of my abilities in this study) and Interest (e.g., I enjoyed participating in this study, This study was a good use of my time). They also proposed several single-item attention and effort items, one of which performed well in their study (i.e., In your honest opinion, should we use your data in our analyses in this study? – with yes/no as the two answer categories). Although self-report measures are easy to administer and have a clear interpretation, their major disadvantage is that it may be unrealistic to expect satisficers to report validly on their own satisficing behavior, especially when the self-report measure uses a format similar to that of the other items in the questionnaire for which the satisficing occurred.

A measure that tries to assess satisficing relatively directly without requiring special scales is based on the time respondents take to complete a survey or a part of a survey (category 2 in Table 3.1). We note that ‘directly’ should be interpreted with caution, since one can never directly measure effort, and time is only an imperfect indicator of effort. As an example of a time-based measure, Wise and Kong (2005) constructed an index labeled response time effort, which is computed as the average of dichotomous scores across all items in a test, where the response to each individual item is scored as 1 when a certain minimum threshold of time taken to respond to the item is exceeded and 0 otherwise. This index was developed for achievement

tests in which there is interest in distinguishing between rapid guessing and solution behavior on individual items. More commonly, response time measures are computed for entire surveys or parts of surveys, such as screens in web-based surveys (Huang *et al*., 2012; Meade and Craig, 2012). The major difficulties with response time measures are that the relationship between response time and effort is unlikely to be linear (low effort is more likely for very short response times), that there are substantial individual differences in the speed of responding (which are likely unrelated to satisficing or optimizing), and that rapid responding does not necessarily signal satisficing, for instance if a pre-existing judgment is readily available (Ferrando and Lorenzo-Seva, 2007).

Given the close relation between attention and visual focus, the time that respondents spend looking at (parts of) an item (i.e., gaze duration) can also be used as an indicator of optimizing (vs. satisficing). In a recent study on reversed, negated and polar opposite items, Baumgartner, Weijters and Pieters (2018) used gaze duration as an indicator of respondent attention. While eye-tracking provides more specific information about respondents' attention compared to overall response times, a major drawback is that it cannot be easily implemented as part of a large-scale online survey, though it should be noted that some basic eye-tracking applications already use web cams.

Several approaches require special scales to detect satisficing but, in contrast to self-ratings of effort, satisficing is assessed via its presumed consequences (category 3 in Table 3.1). One set of techniques is based on items that have a correct response so that unusual answers may be expected to indicate satisficing. Huang *et al*. (2012) discuss this under the label of the "infrequency approach" (e.g., I was born on February 30th), and Meade & Craig (2012) refer to such items as "bogus items" (e.g., All my friends are aliens, see Meade and Craig, 2012, Table 1,

for additional examples). One limitation of these types of items is that they risk upsetting respondent expectations by violating conversational norms (Hauser and Schwarz, 2015; Hauser *et al*., 2016; Schwarz, 1999) and that "asking a strange question may entice people to provide a strange answer" (Baumgartner and Weijters, 2012, p. 566).

A variation on bogus items are instructed response items, where specific instructions to choose a certain response option are given and deviations from the requested answer are regarded as a sign of carelessness (e.g., For this item, do not click on any of the response options; simply leave the item blank). Oppenheimer, Meyvis, and Davidenko (2009) call these items instructional manipulation checks (IMCs). IMCs provide a clear decision rule on whether or not to retain a respondent, and Oppenheimer *et al*. (2009) demonstrate that excluding respondents who failed an IMC can increase the power of statistical tests (see also Kam and Chan, 2018). However, repeated use of IMCs can be annoying to respondents, so they should be implemented sparsely.

In lengthy personality inventories, researchers sometimes include special scales designed to flag inconsistent responders. This is similar to the techniques described below under category 4 measures, except that special scales are used, and this approach will therefore be described in the following section.

The final set of methods used to assess satisficing (category 4 in Table 3.1) attempts to derive satisficing measures from the substantive items in the questionnaire rather than special scales. In addition, satisficing is not measured directly but inferred from response behavior that is presumably due to respondents' reluctance or inability to expend the required cognitive resources. Included in this category are techniques for identifying unusual observations such as outlier detection based on the Mahalanobis D statistic (Meade and Craig, 2012) and person-fit

statistics developed in the item response theory literature (Conijn, Emons, and van Assen, 2013; Emons, Sijtsma, and Meijer, 2005; Meijer, 2003; Reise and Widaman, 1999). Although these methods may be useful in identifying aberrant response behavior, it is doubtful that such behavior is necessarily due to satisficing. In this vein, Sterba and Pek (2012) suggest that person-fit statistics may indicate that a case is problematic, but not necessarily why it is.

The methods that are more specifically designed to identify satisficers based on their responses to substantive scales fall into two classes. On the one hand, responses to similar items should not be too inconsistent (Huang *et al*., 2012; Johnson, 2005; Meade and Craig, 2012). Although specially designed scales could be used for this purpose (as mentioned previously), researchers typically identify items in the questionnaire that are strongly positively or negatively correlated, and then use these items to detect respondents whose response behavior does not show the expected pattern of positive or negative correlations. Unfortunately, the use of inconsistency-based measures of satisficing has some limitations. One problem is that demonstrations of the positive effects of removing inconsistent responders on data quality are somewhat tautological when the inconsistency measure is too closely related to the measure of accuracy used (e.g., internal consistency, clarity of the factor structure). Furthermore, there are sources of inconsistency that are unrelated to satisficing. For example, inconsistency is sometimes assessed via responses to regular and reversed item pairs. As reviewed in Weijters and Baumgartner (2012), misresponding to reversed items has many possible causes and satisficing is only one of them. For example, one important alternative explanation for inconsistent responses are response styles.

On the other hand, responses to questionnaire items (particularly if the items are heterogeneous in content) should not be too similar, because lack of differentiation is likely to

signal nonresponsiveness to item content. For example, Johnson (2005) studied nearly 24,000 protocols of respondents who completed lengthy, web-based personality inventories and found that there was a scree around 9 consecutive identical responses (long strings) for each of the scale positions of a 5-point 'very inaccurate' to 'very accurate' response scale. In other words, if somebody uses the same response option more than 9 times in a row, that person could be regarded as inattentive. Unfortunately, Meade and Craig (2012) found that long strings of the same response correlated only weakly with other measures of satisficing. Long strings of the same response may in part be due to response styles. For example, a respondent high in midpoint response style is more likely to endorse the midpoint repeatedly in response to a series of items. In a sense, then, the long string measure is a suboptimal response style indicator as it focuses on how many adjacent identical responses occur, whereas response style measures are typically not limited to response patterns to adjacent items.

Comparative investigations of the convergent validity of different satisficing measures have shown that the correlations are modest at best. In Huang *et al*. (2012), across two studies a long string measure correlated poorly with a page time and two inconsistency measures (average correlations of .28 and .24, respectively); a self-report measure of effort also showed only modest correlations (.30 and .34, respectively). The average correlation between the page time and the two inconsistency measures was .45. In the study by Meade and Craig (2012), in which a larger set of satisficing measures was used, an exploratory factor analysis yielded a three-factor structure. The bogus items, several inconsistency measures and Mahalanobis D formed one factor; various self-reported effort measures formed another factor; and two long-string measures formed the third factor. The response time factor did not load strongly on any factor.

Several conclusions can be drawn from our review of the satisficing literature. First, some self-report measures of satisficing (category 1) make the questionable assumption that the measures themselves are not subject to satisficing bias. Second, satisficing measures based on exaggerated response inconsistency or consistency (Category 4) may be confounded with response styles (although some response styles can be caused by satisficing). Third, a single measure of satisficing is unlikely to capture the full meaning of satisficing adequately. In sum, most if not all satisficing indicators have their limitations and extant research has yielded few conclusive findings, but until further evidence becomes available two measures of satisficing can be tentatively recommended: (a) the single-item measure "In your honest opinion, should we use your data in our analyses in this study?" (with yes/no as the two answer categories), which was proposed by Meade and Craig (2012); and (b) one or a few instructed response items (of the type "Please select strongly disagree for this item"), as validated by Kam and Chan (2018). While these measures may miss some satisficing respondents (i.e., false negatives), they are unlikely to incorrectly flag many non-satisficing respondents (i.e., false positives). Also, these measures yield clear decisions about whether or not to include a respondent in the study, and they are probably not too annoying to respondents when used cautiously.

3.3 A three-step model of the survey process

Several researchers have developed multi-step models of the process underlying people's responses to survey questions. The most well-known of these is the model proposed by Tourangeau, Rips, and Rasinski (2000), which distinguishes the following five steps: comprehension (interpretation of the question); retrieval (recall of relevant information); judgment (integration of the available information); response mapping (conversion of an internal

judgment into an observable response); and response editing (possible adjustment of the final response). Here, we will use a simplified version of this model, which combines some of the steps because they are difficult to distinguish in practice or because they are not always involved in the response process. The three steps in the revised model are: comprehending what the respondent is being asked to do; formulating a tentative internal response (which often involves some form of retrieval of relevant information from memory); and communicating an overt answer to the researcher (including possibly editing the response).

3.3.1 Comprehension

Most surveys consist of instructions and questions (including the response scale), and respondents have to comprehend both. Instructions are to be understood in a broad sense and may include an acknowledgement of appreciation by the researcher to respondents for their participation in the survey, an explanation of the purpose(s) of the survey, admonitions to take the survey task seriously, introductions to particular tasks to be completed during a survey, and transitions between different parts of a questionnaire. Oppenheimer, Meyvis, and Davidenko (2009) vividly demonstrate that participants do not always read instructions carefully. Participants were shown a screen with the title ‘Sports Participation’ and a question asking ‘Which of these activities do you engage in regularly (click on all that apply)’, followed by buttons listing 10 different sports activities as well as a continue button. However, below the Sports Participation title, instructions in relatively small font told participants to ignore the sports items and to click on the title. In one study, 46 percent of participants nonetheless clicked on the sports categories or the continue button, despite the explicit instructions not to do so. Apparently, participants skipped the instructions. In studies in which it is important that participants pay

careful attention to the instructions (e.g., because the instructions contain a crucial manipulation), lack of attention to the instructions is likely to make the manipulation ineffective.

If respondents are to provide accurate responses, at the very least they must read the questions to which they are asked to respond. Extremely short response time measures and misresponses to instructed response items (in which respondents are asked to click on, say, the strongly disagree response option) or bogus items (discussed earlier) suggest that some respondents do not read the questions that they nonetheless answer. Similarly, inconsistent responses to substantively similar questions for which the polarity of the response scale is varied (e.g., for one question the response scale ranges, say, from 1 = very favorable to 5 = very unfavorable, whereas for another question the range is 1 = very unfavorable to 5 = very favorable) imply that respondents do not always pay sufficient attention to the response scale.

Assuming that respondents do read the question, they have to comprehend both the literal meaning of the question and the implied meaning. Comprehension problems are not necessarily due to respondent inattention, but may instead be caused by poor item formulations. Several researchers have catalogued the most common sources of miscomprehension in surveys. For example, Graesser, Cai, Louwerse, and Daniel (2006) identified five common problems with item wording: unfamiliar technical terms (e.g., myocardial infarction, incongruous), vague or imprecise predicate or relative terms (e.g., often, recently), vague or ambiguous noun phrases (e.g., cultural events, bank), complex syntax (e.g., sentences in which sub-clauses precede the main verb), and working memory overload (e.g., sentences containing multiple ‘or’ or ‘and’ parts). A good example of a problematic item is ‘Quite small setbacks occasionally irritate me too much.’ Research has demonstrated (e.g., based on eye tracking evidence) that problematic

items can have negative effects on various response behaviors (e.g., skipping parts of the question or answering a question before having fully read it).

Building on Graesser *et al*. (2006), Lenzner and colleagues (Lenzner, 2012; Lenzner, Kaczmirek, and Galesic, 2011; Lenzner, Kaczmirek, and Lenzner, 2010) extended the earlier typology of comprehension issues by adding the following two problematic item characteristics: (1) low syntactic redundancy, which refers to a lack of predictability of the grammatical structure, such as passive constructions and/or subordinate sentences (e.g., "Commercials regarding competing brands are not able to reduce my interest in buying the same product (or its successor) again"); and (2) bridging inferences, where respondents have to draw non-obvious inferences to connect different sentences, such as an introduction and the actual question. Lenzner and colleagues also show that problematic items increase the cognitive burden on respondents (as measured by longer response times) and make comprehension more difficult (as assessed by various fixation measures derived from eye tracking evidence). Furthermore, problematic responses can negatively impact response quality (in terms of increased numbers of don't knows, neutral responses, and reduced consistency over time), and the effect may be stronger for respondents lower in ability or motivation.

In a recent study, Hardy and Ford (2014) explicitly asked respondents to explain the meaning of survey questions, using items from several established organizational behavior scales. They distinguished three forms of miscomprehension in surveys – instructional (where respondents do not follow instructions), sentential (where respondents enrich or deplete the original meaning of a sentence), and lexical (where respondents interpret a word differently) – and they showed that all three forms of miscomprehension occurred. As an example of problems resulting from sentential miscomprehension, Hardy and Ford (2014) point out that many

respondents ‘miss’ the process element in a scale that measures procedural justice and hence answer a question about distributive justice, which might help explain the often-high correlation between procedural and distributive justice. As another example, about one in five respondents interpreted the item ‘‘I am satisfied with my job for the time being’’ as ‘‘At the moment I am satisfied with my job and I am not looking for a new one,’’ an interpretation that also taps into turnover intention; thus, the content validity of the satisfaction measure is reduced and its correlation with turnover intention may be overestimated. Overall, they find that half of respondents deviate from the strict syntax of items and alter it according to their own understanding. Based on the foregoing findings, some key recommendations concerning item wording are reported in Table 3.2 (Graesser *et al*., 2006; Graesser, Wiemer-Hastings, Kreuz, Wiemer-Hastings, and Marquis, 2000; Hardy and Ford, 2014; Lenzner, 2012, 2014).

Most researchers probably assume that respondents correctly interpret the response scales associated with a question. This is not always the case. Arce-Ferrer (2006) conducted a study with senior high school students in Mexico in which respondents had to fill in the missing intermediary response category labels on a 7-point Likert scale with endpoints of totally agree and totally disagree (i.e., only the endpoints were labeled). Respondents’ subjective categories frequently deviated from the intended scale categories of moderately agree to moderately disagree, with a midpoint of neither agree nor disagree; examples of problematic interpretations included “it does not bother me”, “all right”, “forget it”, “I liked it”, or “I feel uncomfortable.”

An additional complication is that respondents may draw (possibly unintended) inferences about the question being asked or the response to be provided from the categories of the response scale. For example, “feeling really irritated” is interpreted differently when the response options range from “several times a day” to “less than twice a week” rather than “more

than once every three months" (Schwarz *et al*., 1988). As another example, the endpoint label "not at all successful" may be interpreted differently when numeric values of -5 to +5 are used for the response scale (i.e., failure) instead of 0 to 10 (lack of success) (Schwarz *et al*., 1991).

Finally, respondents sometimes draw inferences about questions based on question context. For example, earlier questions may inform people's responses to later questions, such as when German university students expressed stronger support for an "educational contribution" when the question was preceded by an item about government financial support in Sweden compared to an item about college tuition in the U.S. (Strack *et al*., 1991).

It is often difficult to fully anticipate comprehension problems attributable to either item wording or the use of particular response scales. Even experts are not always able to identify all problems, so there is no substitute for in-depth pretesting (Baumgartner and Weijters, 2012). This includes pilot tests, cognitive interviews, and possibly even eye-tracking for expensive surveys. These methods can be onerous, but if the survey is sufficiently important, the effort spent on thorough pre-testing is time and money well-spent.

3.3.2 Judgment

Although respondents may be asked many different types of questions, most of the time the questions are closed-ended (rather than open-ended) and the respondent is required to render some kind of overall judgment. It is thus important to understand how judgments are formed and what factors influence judgment formation.

We will distinguish three prototypical judgment tasks, depending on whether a judgment is memory-based (involving the retrieval of information from long-term memory) or made on-line (i.e., based on information externally present in the judgment context), and whether or not a

previously formed judgment is already available in long-term memory (see Hastie and Park, 1986): (a) information present in the external environment is used to form an on-line judgment; (b) a previously formed judgment is retrieved from long-term memory; and (c) information available in long-term memory is retrieved and used to render a judgment. A fourth possibility is a combination of (a) and (c), but this situation need not be considered separately.

Case (a) represents situations in which the information necessary to form a judgment is (mostly) externally available in the stimulus environment (although interpretation of external information also requires information stored in memory). An example is a product evaluation or choice task in which an unknown product (several unknown products) is (are) described on various attributes and the respondent is asked to provide an overall evaluation of the product or choose one of the products (as in typical conjoint studies). Much research has investigated how consumers integrate attribute information into overall evaluations (Wilkie and Pessemier, 1973) and which choice rules they use when selecting products (Bettman, 1979). More recently, the literature has emphasized that evaluations and choices are often highly contingent on a multitude of factors related to the respondent, the task, and the context (Bettman, Luce, and Payne, 1998), including factors that depend on how preferences are elicited (e.g., preferences may be reversed in evaluation and choice tasks). In general, judgments are probably much more labile than is often acknowledged because information that happens to be temporarily accessible in the judgment situation can have a strong influence on people's answers.

Case (b) is representative of situations in which respondents answer questions that they have encountered many times, which implies that the answers are well-rehearsed. Responses to demographic questions fall into this category, and there is little reason to believe that the retrieval of previously formed judgments of this kind is problematic. Of course, there are other

examples in which previously formed judgments are already available in long-term memory (e.g., when beer aficionados are asked to evaluate one of their favorite beers), but in general instances in which respondents can simply retrieve a required judgment from long-term memory are probably rare.

Case (c) is probably the most common situation in many real-world survey settings. Respondents usually do not have ready answers available for all the things they might be asked about, and they thus have to construct an answer on the spot. Information that happens to be available in the survey setting may influence the answers (as already mentioned), but usually the task requires the retrieval of relevant information from long-term memory. As discussed earlier, it is unlikely that respondents routinely engage in an extensive memory search prior to rendering a judgment (Krosnick, 1991), and there are well-known shortcomings of human memory that can make the retrieval of stored information problematic. Schacter (1999) discusses seven sins of memory, six of which are highly relevant for survey research. These include three sins of omission or types of forgetting (decreasing accessibility of information over time; poor encoding or retrieval of information; and temporary inaccessibility of stored information) and three sins of commission or memory distortions (misattribution to an incorrect source, or false recognition and recall; false recollections due to leading questions; and distortion of the past by knowledge of the present). Apart from the fact that respondents have to be encouraged to engage in memory search, various strategies can be used to counter these sins of memory, such as providing helpful retrieval cues or avoiding leading questions.

One important source of error at the judgment stage relevant to case (c) is confirmation bias, which is similar to the issue of leading questions but more general. Confirmation bias refers to the phenomenon that respondents tend to retrieve information from memory that supports the

question being asked (Davies, 2003; Kunda *et al*., 1993; Weijters, Baumgartner, and Schillewaert, 2013). For example, when respondents are asked whether they are extraverted (introverted), they will think of situations in which they were extraverted (introverted). Thus, respondents asked about extraversion will likely rate themselves as higher in extraversion than those asked about introversion, and vice versa. Since there is evidence that the bias is stronger when the judgment requires a search for relevant information in memory (because an overall judgment is not available) and when respondents can retrieve sufficient information consistent with the way the question is being asked, care has to be taken to avoid the bias when these conditions apply. One way of doing so would be by asking two-sided questions such as "Do you agree or disagree with the following statements?" An even better method might be to present the core proposition in a two-sided fashion by using item-specific response options, for instance: "How would you rate your health – excellent, very good, fair, or bad?" (Saris, Revilla, Krosnick, & Shaeffer, 2010).

Another important issue that must be considered carefully is the positioning of items in a survey. Studies have demonstrated that responses to items that are located in close proximity are correlated more strongly than responses to items that are positioned farther apart (note that this so-called proximity effect reverses for reversed items; Weijters *et al*., 2009). Various explanations for this finding have been suggested, including (local variations in) response styles such as acquiescence (Hui and Triandis, 1985; Weijters, Geuens, and Schillewaert, 2010a), anchoring and adjustment (Gehlbach and Barge, 2012), and cognitive carry-over (Harrison and McLaughlin, 1993). Different conclusions for questionnaire design have been drawn from these findings. Some authors recommend that items measuring the same or similar constructs should be grouped together, because it places less cognitive burden on respondents and enhances

internal consistency within constructs and discriminant validity across constructs (Harrison and McLaughlin, 1996). Other authors argue that better convergent and discriminant validity caused by blocking items may be a methodological artifact (Weijters *et al*., 2014) and that varying the keying direction of items from the same scale and mixing items from different scales (item dispersal or randomization) leads to better coverage of the full conceptual domain of a construct, even if internal consistency may suffer somewhat.

A final judgmental bias that can negatively affect the validity of measurement is (illusory) halo bias (Cooper, 1981; Lance, LaPointe, and Fisicaro, 1994). This occurs when a respondent's general impression of the focal stimulus influences his or her dimensional ratings, or when ratings on a salient dimension affect ratings on other dimensions. Halo bias leads to exaggerated correlations between dimensional ratings, making it difficult to uncover the dimensional structure underlying perceptions of the focal stimulus and creating methodological problems such as multicollinearity.

3.3.3 Response

We will focus on situations in which respondents have to map their internal response onto a numerical response scale (usually a closed-ended question with a fixed number of response options). Typical examples are agree-disagree scales, semantic differential scales of all kinds, and binary and multiple-choice questions. One phenomenon that complicates things at the response mapping stage is that the choice of a response option is not only determined by substantive considerations but can also be affected by content-irrelevant factors. A common source of bias are so-called response styles, which are systematic preferences for certain response options on rating scales (Baumgartner and Steenkamp, 2001). They include

acquiescence response style (a preference for the agreement options or, more generally, the positive side of rating scales), disacquiescence response style (a preference for the disagreement options or, more generally, the negative side of rating scales), extreme response style (a preference for the most extreme options on either side of the rating scale), and midpoint response style (a preference for the middle or neutral position of the rating scale). In the extreme, these response influences occur independently of the substantive content of the items, thus constituting a particularly serious distortion of reality, although in practice it seems unlikely that respondents will ignore substantive considerations completely.

Response styles depend on various situational factors, including features of the items to which people respond, as well as dispositional characteristics of respondents. Table 3.3 provides an overview of the conceptualization and measurement of the most common response styles (based on Baumgartner and Steenkamp, 2001). Research has also demonstrated that there are reliable differences in response styles across cultures (see Baumgartner and Weijters, 2015, for details), which implies that cross-cultural researchers have to be very careful when they want to conduct cross-cultural comparisons in the presence of differential response styles.

Response styles are problematic because they add extraneous variability to observed measurements. Furthermore, the resulting measurement error is usually systematic, which implies that both means on substantive variables and relationships between variables can be seriously distorted. It is thus important to prevent or control the operation of response styles. With regard to pre-data collection techniques aimed at preventing stylistic responding, the major recommendation is to encourage systematic processing, since peripheral processing conditions (low respondent ability and motivation) are generally associated with all forms of stylistic responding. There are also some remedies for specific types of response styles, such as the use of

balanced scales (i.e., scales with an equal number of regular and reversed items) to discourage acquiescent responding.

If stylistic responding is a serious problem, one can also try to control for the problem statistically after the fact. Several approaches have been considered for this purpose. For example, Böckenholt (2017) proposed so-called item response tree models, which can be used to decompose the overall response to a rating scale into an ordered sequence of queries such as whether or not to choose the midpoint of the rating scale, whether to indicate weak or strong (dis)agreement with an item, and whether or not to select one of the most extreme scale positions. Researchers who only want to reduce bias due to midpoint and/or extreme response styles can recode the original responses to binary agree/disagree variables and use these binary indicators in subsequent analyses using Item Response Theory (IRT) models (Zettler, Lang, Hülsheger, and Hilbig, 2015). However, this procedure results in a loss of information and is valid only under the assumption that agreement is substantive, whereas midpoint and extreme responding are stylistic in nature. Other advanced IRT-based modeling approaches can be used to account for response styles by incorporating them into measurement parameters, for instance by modeling separate IRT parameters for different individuals (Bolt, Lu, and Kim, 2014), for different observed groups such as respondents from different countries (De Jong, Steenkamp, and Fox, 2007), or for different latent classes (Morren, Gelissen, and Vermunt, 2011).

Alternatively, one or multiple response style measures can be included as covariates (Weijters, Schillewaert, and Geuens, 2008) or used to recode response options in a group-specific way to model between-group response style differences (Weijters, Baumgartner, and Geuens, 2016). Unfortunately, good response style measures generally require scales whose only purpose is to assess stylistic responding (i.e., they serve no substantive purpose), so researchers

are generally reluctant to include such measures since they increase the burden imposed on respondents. The two main types of valid response style measures that control for content are measures based on response patterns across many heterogeneous items (Baumgartner and Steenkamp, 2001; Greenleaf, 1992a, 1992b; Weijters *et al*., 2008) and measures based on responses to questions that presumably have a known true value, that is, anchoring vignettes (Bolt *et al*., 2014; King and Wand, 2007), although we are not aware of marketing research that has used anchoring vignettes. Even though researchers have tried to construct response style measures from the substantive measures themselves, it is unrealistic to assume that content and style can be cleanly separated when the same items are used to measure both (except for special circumstances, such as when reversed items are available). Apart from response styles, there are also response tendencies where people do not ignore the content of the items, but the observed response does not reflect the true response. This is sometimes called a response set, in contrast to response styles. The most well-known example of a response set is socially desirable responding (SDR), which was already discussed earlier.

Response styles and response sets are usually treated as relatively stable individual difference variables that affect people's response behavior (Weijters, Geuens, and Schillewaert, 2010a, 2010b). However, there are many other characteristics of response scales that can influence how respondents map an internal response onto the response scale provided in the questionnaire. In what follows, we discuss the issues surrounding the number of response options and the way they are labeled, and then address a few other topics related to the format of the response scale.

Deciding on the number of response options to use in a rating scale can be divided into two questions: (1) whether or not to include a midpoint (i.e., whether to use an odd or even

number of response categories), and (2) how many response categories to employ, not counting the midpoint.

The issue of whether or not to include a midpoint has been debated for decades. Opponents argue that the midpoint provides respondents with a ready opportunity to avoid thinking about the issue under investigation, which suggests that omitting the midpoint may improve data quality (Converse and Presser, 1986). Those in favor of offering a middle category point out that in its absence, respondents who do not have knowledge on the subject will choose one side or the other, which increases the error in survey data (Nadler, Weston, and Voyles, 2015; Nowlis, Kahn, and Dhar, 2002). Not offering a midpoint forces respondents to take a stance, but this can sometimes be problematic. Respondents who are truly neutral are forced to select a response that is not reflective of their true opinion. This results in increased random error and/or systematic bias. Random error reduces reliability and weakens correlations with other constructs; systematic bias can cause spurious effects. The lack of a midpoint option may also irritate respondents. Ambivalent respondents who are forced to take a position tend to react negatively (Nowlis *et al*., 2002; Weijters, Cabooter, and Schillewaert, 2010). If choosing the midpoint always indicated a neutral opinion, or a balance of reasons favoring either agreement or disagreement (i.e., attitudinal ambivalence), the recommendation to use a midpoint option would be noncontroversial. Unfortunately, the neutral response may also be chosen for other reasons. Respondents interpret the midpoint in different ways and they cite different justifications for choosing the midpoint (Baka, Figgou, and Triga, 2012; Nadler *et al*., 2015), including lack of an opinion or knowledge about an issue (i.e., don't know or DK, which is sometimes used as a separate response option); indecision or uncertainty about one's opinion; indifference or lack of interest; evasiveness or a desire not to reveal one's true opinion (e.g., for reasons of social

desirability); an attempt to dispute aspects of the question; confusion about the question; and incorrect interpretation of the midpoint (e.g., the midpoint is seen as slight agreement or slight disagreement). These reasons make the midpoint an ambiguous repository of different meanings.

Our recommendation is to include a midpoint option, but to ensure that respondents' choice of the midpoint accurately reflects a neutral opinion, or at least an ambivalent stance. This means that items have to be formulated clearly and unambiguously, that respondents must feel comfortable answering the question and do not feel compelled to hide their true opinion, and that respondents are presented with questions that they know about and that are relevant to them (e.g., filter questions can be asked to determine respondent eligibility for follow-up items), so they will take the time to form an opinion even when none previously existed. If these conditions are met, the meaning of a midpoint response should be relatively unambiguous and including a midpoint option can be advantageous, because it also reduces inconsistent responding to reversed items (Weijters *et al.*, 2010).

Once a decision has been made about whether or not to include a midpoint, the researcher has to decide how many response options (apart from the midpoint) to use. When addressing this question, researchers face a tradeoff between information richness and interpretability. On the one hand, from an information theory perspective, a scale range must be refined enough to allow for maximal information transmission (Cox III, 1980; Garner, 1960; Green and Rao, 1970). This suggests that too few scale steps should be avoided. On the other hand, from a respondent perspective, increasing the number of categories makes the response task more difficult and may entail a level of precision that is no longer meaningful. This suggests that too many scale steps should be avoided. From a statistical perspective, analyzing categorical data as if they were continuous has been shown to be acceptable when using five or more response categories (Bollen

and Barb, 1981; Srinivasan and Basu, 1989). From a respondent perspective, formats with a small number of response categories (e.g., fewer than four) are evaluated as quick to use but also as poor for adequately expressing one's feelings (Preston and Colman, 2000). Based on survey experiments and simulations in which they evaluated the effect of various scale characteristics on response biases, Weijters *et al*. (2010) recommend five-point scales for surveys among the general population and five- or seven-point scales for surveys among more experienced respondents (e.g., students, MTurkers). In sum, combining both the information-theory and respondent perspectives, five- and seven-point response formats can be recommended because they yield an adequate tradeoff between the loss of information entailed by fewer scale steps and the complexity of the judgment task implied by more response categories. The key advantage of five-point scale formats lies in the unambiguous interpretation of the response categories, and five categories seem sufficiently fine-grained for common statistical analyses based on the general linear model.

Closely related to the choice of the number of response categories is the labelling of the categories. An important function of labeling is to disambiguate the meaning of the response categories. Questionnaires often employ rating scales that have verbal labels attached to the endpoints of the response scale only. Presumably, such scales are easier to design and may intuitively be better aligned with the common assumption that responses are measured on an interval scale (i.e., when consecutive numbers are used for the scale steps, an interval level of measurement is implied). However, research shows that formats with verbal labels for all categories facilitate interpretation and enhance reliability (Krosnick and Fabrigar, 1997; Wildt and Mazis, 1978). If only the endpoints are labeled, the non-labeled categories may be hard to interpret for some respondents, as evidenced by the ambiguous and divergent interpretations that

respondents provide when asked to write down the meanings of response categories (Arce-Ferrer, 2006). In contrast, when all scale positions are fully labeled, all categories are more or less equally clear to respondents (Cabooter, Weijters, Geuens, and Vermeir, 2016; Moors *et al.*, 2014). Weng (2004) demonstrated that rating scales with clear labels for all the response options result in higher test-retest reliability than scales in which only the endpoints are labeled. Recently, Moors *et al*. (2014) demonstrated that labeling only the endpoints evokes more extreme response style bias than full labeling. Weijters *et al*. (2010) found that five-point formats with labels for all five categories resulted in more consistent responses to reversed items than the other response formats tested in their study (which ranged from four to seven response categories). This suggests that the response categories are least ambiguous in this format. In summary, the empirical evidence supports the recommendation that all scale positions should be labeled.

In a recent paper, DeCastellarnau (2018) proposed a classification of response scale characteristics and reviewed prior empirical evidence relevant to each of these characteristics. Her summary of the findings concerning scale characteristics that have been found to reliably influence data quality can be summarized as follows: (1) agree-disagree scales (e.g., I am satisfied with this product, rated on an agree-disagree scale) have lower data quality than direct ratings of the dimension of interest (e.g., Please rate your satisfaction with this product on a scale ranging from dissatisfied to satisfied); (2) type of scale (DeCastellarnau distinguishes four types of continuous scales, such as scales for which respondents have to input a number or scales on which respondents have to mark a point on a continuum, and four types of categorical scales, such as dichotomous scales and rating scales with three or more response categories) has an effect on data quality, but the effects are complicated and there are few general conclusions; (3)

the length of a scale (e.g., the number of response categories in the case of categorical scales) matters, as discussed previously; (4) the labeling of the response categories influences data quality, as already mentioned; (5) the use of so-called fixed reference points (always, never, completely, etc.) increases measurement quality because it enhances the comparability of measurements across people; (6) (the order of) numerical labels can influence response distributions (e.g., negative to positive, positive to negative, 0 to positive, 0 to negative, etc.), but no general recommendations can be provided; (7) numeric labels should correspond with verbal labels (e.g., -5 to +5 for a bipolar bad to good scale, or 0 to 10 for a not at all to completely unipolar scale); (8) the presence of a neutral response alternative impacts data quality, as discussed before; (9) graphical rating scales (e.g., ladders, thermometers, dials, etc.) have an effect on data quality, but the findings are difficult to interpret; (10) the layout display of scales (e.g., horizonal vs. vertical) affects responses, although there seem to be no general conclusions; and (11) the visual separation of labels (i.e., some or all of the response options are separated, for example by putting them in boxes) seems to affect data quality (e.g., separation may decrease nonresponse and improve reliability).

In summary, when planning a survey, researchers should evaluate the survey task from the respondent's perspective. First, respondents do not always complete a survey with an accuracy goal, and if goals other than accuracy are likely to be salient, the accuracy goal has to be reinforced, and it may be necessary to explicitly measure the presence of other goals that have the potential to distort the findings of a survey (e.g., social desirability). Second, respondents are frequently far less involved in surveys than the researcher, and sometimes they lack the ability to respond accurately. If satisficing behavior is likely, special efforts have to be made to encourage optimizing, and usually it is a good idea to measure satisficing explicitly so that controls for

satisficing behavior can be enacted after the fact. Third, answering survey questions is frequently not a straightforward process, and a respondent has to go through several steps to render a response, at each of which errors may occur. Researchers have to think carefully about these sources of error and design their surveys so that mistakes can be minimized. Careful pretesting of surveys is probably the most effective strategy for ensuring accurate results.

4

Measurement as the formulation of measurement models linking observable indicators to latent concepts

Sometimes, a single observed measure is used to capture a theoretical concept, which assumes that there is no measurement error of any kind in the observed response variable. Rossiter (2002) argues that for "completely concrete constructs, one concrete item is all that is necessary" (p. 321). A "completely concrete construct" is one for which both the object of measurement and the attribute to be measured are concrete. Bergkvist and Rossiter (2007) expand on this and propose that "a single-item measure is sufficient if the construct is such that in the minds of raters (e.g., respondents in a survey), (1) the object of the construct is 'concrete singular,' meaning that it consists of one object that is easily and uniformly imagined, and (2) the attribute of the construct is 'concrete,' again meaning that it is easily and uniformly imagined" (p. 176). As an example of such a construct they mention attitude toward the ad (or ad liking, A_{ad}) and attitude toward the brand (brand attitude, A_{brand}), and they report a study in which they compare the predictive validity of single- and multi-item scales. Specifically, they showed a small sample of students (n=92) real ads for four different real products (painkillers, coffee, pension funds, and jeans), which were not available in the local market and thus new to respondents, and asked respondents to rate both the ad and the brand on multiple scales (one ad liking measure and three semantic differential measures each of A_{ad} and A_{brand}, such as good-bad). Correlations between either a single-item or multi-item measure of A_{brand} and a single- or multi-item measure of A_{ad} (including ad liking) indicated that multi-item scales (of either the independent or dependent variable) did not lead to significantly higher correlations than single-item scales. Bergkvist and Rossiter

(2007) thus conclude that “for the many constructs in marketing that consist of a concrete singular object and a concrete attribute, such as A_{ad} or A_{brand}, single-item measures should be used” (p. 175).

It should be clear that the conclusions of Bergkvist and Rossiter (2007) have limited applicability (despite the reference to “the many constructs in marketing”) because most constructs are not doubly concrete, especially in research that is motivated by theoretical concerns (i.e., most of the research found in academic journals). What’s more, the recommendations of Bergkvist and Rossiter (2007) should be treated with caution even when constructs are doubly concrete, as argued in a recent paper by Kamakura (2015). He points out that it is impossible to assess reliability and, if necessary, correct for random measurement error when multi-item measures are unavailable. The findings of Bergkvist and Rossiter (2007) are based on very limited evidence, and it seems foolhardy to accept their conclusions without being able to ascertain whether they apply in a given context. Furthermore, Bergkvist and Rossiter (2007) never applied a correction for attenuation, and their comparison of correlations between single- and multi-item measures was thus incomplete. Finally, Kamakura (2015) notes that Bergkvist and Rossiter (2007) did not assess predictive validity but concurrent validity, and that common method bias cannot be ruled out convincingly in their study. Kamakura (2015) reports a “true” predictive validity study in which self-reported attitudes toward weight and natural food are correlated with later purchases of low-fat and organic milk, and he shows that multi-item attitudinal measures corrected for attenuation yield consistently higher correlations with purchase behavior than single-item measures (although the differences in correlations are admittedly small).

Although most researchers would agree that multiple items should be used economically, as a general rule of thumb constructs should not be measured with single items. Even if single items turn out to be reasonably reliable and valid in a given situation, the researcher does not know whether a single item can be trusted unless multi-item measures are available. In the sequel, we will therefore assume that researchers use multiple indicators to measure their constructs in an effort to more faithfully capture the theoretical entities of interest. This is necessary even by Rossiter's standards when constructs are *not* doubly concrete (which is usually the case), and it is the safer option even when constructs *are* doubly concrete. The question then becomes how the correspondence between the observed measures and their intended constructs can be ascertained.

Even when multiple measures of a construct are available, researchers frequently collapse them into a single overall composite (by summing or averaging the individual measures) or combine subsets of items into parcels of items (particularly when the number of individual items is large; see Table 2.1 for examples). Averaging individual items will usually result in more reliable and valid assessments of the intended construct (compared to single-item measures), but it should only be done for well-validated scales or after a careful measurement analysis has been conducted. Reporting a coefficient alpha estimate that's "sufficiently high" by (arbitrary) conventional standards (e.g., greater than, say, .7) does not provide convincing justification for averaging and is not an adequate substitute for a thorough measurement analysis. In general, it is best to take into account unreliability of measurement directly when relating the construct in question to other constructs (by using an explicit measurement model), but if there are too many items, it is possible to correct for measurement error by (a) using an average of the available items as a single indicator of the underlying construct (after unidimensionality of the items in the

composite has been established), (b) fixing the factor loading to one, (c) setting the unique variance to one minus the reliability of the composite (e.g., based on coefficient alpha) multiplied by the variance of the composite, and (d) freely estimating the factor variance. Alternatively, a researcher may form item parcels based on the available indicators (e.g., an 18-item scale may be split into three parcels of six items each). For unidimensional scales, the assignment of items to parcels can be done randomly, whereas for multidimensional scales two parceling methods can be considered (Cole, Perkins, and Zelkowitz, 2016). One is homogeneous parceling, in which the items that are combined into a parcel represent a single lower-order dimension of the (higher-order) construct. The other is heterogeneous parceling, in which the items that are combined into a parcel represent all lower-order dimensions of the (higher-order) construct. Based on simulated and real data sets, Cole *et al*. (2016) find that both approaches can result in models that fit the data well. They also show that, compared with homogeneous parceling, heterogeneous parceling generates smaller (i.e., closer to zero) but tighter estimates of structural path coefficients, the net result of which is greater statistical power to test substantive relations among latent variables. We note that homogeneous parcels can be argued to better represent the multiple dimensions of the construct, provided the dimensions are reasonably highly correlated (otherwise the parcels won't cohere).

Even within a given parceling strategy (e.g., heterogeneous parceling), items can be assigned to parcels in many different ways. The way in which items are assigned to parcels (so-called parcel allocations) has been found to affect both parameter estimates and model fit results, and this phenomenon is called parcel-allocation variability (Sterba, 2011; Sterba and Rights, 2017). To account for parcel-allocation variability, researchers can create multiple datasets with different parcel allocations to empirically quantify its effects (SAS and R syntax to automate this

process is provided in Sterba, 2011, and Sterba and Rights, 2017). Generally speaking, item parceling should not be used for scales whose factor structure is not well-understood, or when a researcher wants to assess the invariance of measurement across populations (e.g., different countries). Holt (2004), Bandalos (2002) and Cole *et al*. (2016) provide additional detail.

As pointed out by MacKenzie *et al*. (2005), a crucial distinction between measurement models centers on the flow of causality between indicators and constructs. If causality flows from the construct to the indicators, the measurement model is reflective and the indicators are called reflective (or effect) indicators; if the causality flows from the indicators to the construct, the measurement model is formative and the indicators are called formative (or cause) indicators. MacKenzie *et al*. argue that if (a) indicators are manifestations of an underlying construct rather than defining characteristics of the construct in question, (b) any one indicator is conceptually interchangeable with the other indicators of the same construct, (c) indicators will necessarily covary, and (d) each indicator has the same antecedents and consequences as the other indicators of the same construct, then the measurement model is reflective; otherwise it is formative.

Jarvis *et al*. (2003) reviewed a large number of measurement models reported in four leading marketing journals and found that the direction of causality between indicators and constructs was often misspecified (in 29% of the cases studied). Most often the misspecification was due to formative measures being treated as reflective measures (see also Diamantopoulos *et al*., 2008, for a summary of other studies presenting evidence of measurement model misspecification). In the opinion of the present authors, formative measurement models are problematic in many ways (see the discussion below, as well as Edwards, 2011, and Howell, Breivik, and Wilcox, 2007), but it is undeniable that reflective measurement models are not

always appropriate. Below we will discuss both reflective and formative measurement models, although our focus will be on the former since they are more common.

A second important distinction related to measurement models deals with whether the observed variables are (assumed to be) discrete or continuous, and whether the scale (or level) of measurement is (assumed to be) nominal, ordinal, interval, or ratio. Conceptually, a variable is discrete (continuous) if its potential values can (cannot) be counted. In practice, a variable is treated as discrete (continuous) if the number of distinct values is relatively small (large), although continuous variables are sometimes discretized (e.g., age is categorized into different age groups). Even when there are relatively few distinct values, researchers often treat variables as continuous. For nominal variables, numbers simply denote group membership; for ordinal variables, numbers denote rank-order (greater or smaller than); for interval variables, differences between numbers are meaningful (i.e., if the difference between two pairs of numbers is the same, the distance between them is the same); and for ratio scales, ratios are meaningful because an absolute zero exists (e.g., one can say that a number is twice as large as another). Interval and ratio scales are sometimes referred to as metric scales and we will not distinguish between the two in the sequel. Table 4.1 presents a cross-classification of variables based on continuity and level of measurement and lists illustrative examples.

In this monograph we will not deal with measurement models for variables measured on a nominal scale. We will start with a discussion of reflective measurement models that are designed for continuous observed variables measured on a metric scale. Strictly speaking, this limitation makes these models inapplicable to most of the observed variables encountered in empirical research in marketing, because most variables are neither continuous nor metric. Nonetheless, researchers routinely apply metric measurement models developed for continuous

data in the hope that the findings will be reasonably robust to violations of this assumption. Some evidence supporting this hypothesis will be presented below, particularly when robust estimation methods that correct for violations of normality of the data are employed. Nonetheless, we will also discuss models specifically designed for discrete ordinal variables, which are the data usually available for analysis. These models are more complex than the models that are usually used, but software to estimate them is becoming more readily available and measurement models for discrete ordinal data deserve more widespread use. We will not deal with discrete ordinal measurement models in the context of formative indicator models because, on the one hand, the basic issues involved are similar to those encountered with reflective indicator models, and, on the other hand, the complexities associated with modeling discrete ordinal data would further exacerbate the many problems afflicting formative indicator models.

The measurement models discussed below play an important role in the construct validation process. Viewed broadly, construct validity incorporates three types of considerations: (a) the theoretical meaningfulness of the construct by itself; (b) the correspondence between the construct and its empirical measures; and (c) the place of the construct within a nomological net of related constructs as stipulated by some theory (e.g., Bagozzi, 1980). The first consideration only deals with the conceptual domain and is thus not directly relevant to measurement, although a clear conceptualization of the construct is a prerequisite for measurement and various issues relevant to construct conceptualization were discussed in section 2. The third consideration is also situated in the conceptual domain, but an assessment of nomological validity requires a theory that imposes a structure on the constructs represented in the nomological net and enables the derivation of propositions about how the focal construct is related to various antecedents and

consequences, or how these relationships might be moderated by additional constructs. Again, since these issues do not involve measurement directly, they will not be discussed here, although observed measures of all constructs represented in the theory have to be available if the nomological net is to be examined empirically. The second consideration connects the conceptual and empirical domains and is thus most relevant for our discussion of measurement. Some important issues related to the correspondence between constructs and measures were already discussed in section 2. However, other important issues will be covered below. Specifically, a careful measurement analysis will provide evidence about the following issues, which are generally regarded as key requirements of construct validity: unidimensionality of the multiple measures (indicators) of the construct overall or within each subdimension of the construct; reliability and convergent validity of the indicators; discriminant validity of the construct or its measures from related constructs and their measures; and invariance of measurements across persons, settings, and times (Steenkamp and van Trijp, 1991). One of the most influential approaches to construct validation has been the multitrait-multimethod (MTMM) technique introduced by Campbell and Fisk (1959), which was proposed as a way of testing convergent and discriminant validity. An updated version of this method based on confirmatory factor analysis will be discussed below, but other approaches for investigating the presence of method effects will be described as well.

4.1 Reflective measurement models

The most common type of reflective measurement model is the confirmatory factor analysis (CFA) model assuming continuous and metric observed (manifest) measures (indicators or items). Although observed variables are usually not continuous, researchers generally assume (or

hope) that if an observed variable can take on at least 5 distinct values, the continuity (or quasi-continuity) assumption will be reasonable. Similarly, even when variables are not measured on a metric scale, the hope is that violations of this assumption will not invalidate the results. However, measurement models for ordered categorical data are available and are becoming somewhat more common. We will consider both types of models in this section.

4.1.1 The basic reflective measurement model for continuous metric observed variables

A reflective measurement model for continuous observed measures measured on a metric (interval or ratio) scale assumes that each (mean-centered) observed variable x_i is a linear function of a (mean-centered) common factor ξ_j and a unique factor δ_i:

$$x_i = \lambda_{ij}\xi_j + \delta_i \tag{4.1}$$

The common factor captures sources of variability that are common to all observed measures of the same ξ_j, and it is supposed to represent the construct of interest, although this need not be the case (i.e., there is no guarantee that the common part of the observed measures actually captures the construct of interest, and only the construct or interest). In the sequel we will use construct and common factor synonymously, although this caveat should be kept in mind. The strength of the relationship between x_i and ξ_j is measured by λ_{ij}, which is called a factor loading. The unique factor δ_i represents all influences on the observed measure other than the variability common to all observed measures of the presumed underlying construct. Frequently, it is assumed that δ_i models random measurement error, although this is a heroic assumption in most cases. An important issue is how the unique factors of different items measuring the same construct are related. The usual assumption is that the unique factors of different items are uncorrelated. This implies that the only source of covariation between items measuring the same construct is the

presumed underlying construct. If the unique factors represent random measurement error, this is a reasonable assumption. In practice, it is unlikely that the unique factors corresponding to different items measuring the same construct are uncorrelated (e.g., because of the way the items are worded).

When multiple common factors are included in the factor model, the model specification must state (a) how many common factors there are, (b) how the observed variables are related to the various common factors, and (c) how the unique factors corresponding to indicators of different common factors are related to each other. In studies in which measurements are collected to measure specific constructs or dimensions of constructs, the number of common factors is generally specified *a priori*, although the model specification is sometimes revised based on the empirical findings. Generally, observed indicators are hypothesized to measure one and only one construct (the so-called target construct) and loadings on non-target constructs are hypothesized to be zero. Again, this is often an unrealistic assumption. For example, if the item ‘‘I am satisfied with my job for the time being’’ is interpreted as ‘‘At the moment I am satisfied with my job and I am not looking for a new one,’’ this item measures not only job satisfaction but also turnover intention (see Hardy and Ford, 2014). The unique factors of indicators of different constructs are routinely specified to be uncorrelated, but often this is an unrealistic assumption. It is possible to relax some of the forgoing assumptions (e.g., non-target loadings need not be specified to be zero, and unique factors can be allowed to be correlated), but this is rarely done in practice. An exception occurs when common method variance is modeled using a method factor or correlated uniquenesses (which can also be done for indicators of the same construct).

In matrix form, the reflective measurement model for continuous and metric observed variables can be stated as

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{4.2}$$

where $\mathbf{x}$ is an I x 1 vector of observed measures, $\boldsymbol{\Lambda}$ is an I x J matrix of factor loadings, $\boldsymbol{\xi}$ is a J x 1 vector of common factors, and $\boldsymbol{\delta}$ is an I x 1 vector of unique factors. Assuming that $\mathbf{x}$ and $\boldsymbol{\xi}$ are in deviation form (i.e., mean-centered), that the expected value of $\boldsymbol{\delta}$ is zero (i.e., $E(\boldsymbol{\delta}) = \mathbf{0}$), and that $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$ are uncorrelated (i.e., $Cov\ (\boldsymbol{\xi}, \boldsymbol{\delta}') = \mathbf{0}$), the variance-covariance matrix of $\mathbf{x}$ (which is called $\boldsymbol{\Sigma}$) is given by:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{4.3}$$

where $\boldsymbol{\Phi}$ (with typical element φ_{ij}) and $\boldsymbol{\Theta}$ (with typical element θ_{ij}) are the variance-covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$, respectively, and the symbol $'$ is the transpose operator.

As discussed previously, researchers generally assume that each observed variable loads on a single common factor (i.e., $\boldsymbol{\Lambda}$ contains only one nonzero entry per row) and that the unique factors are uncorrelated (i.e., $\boldsymbol{\Theta}$ is diagonal). The resulting model is called a congeneric measurement model. More restricted versions of this model are obtained when all the factor loadings for a given construct are restricted to be the same, which implies that the scale metrics of the indicators are identical (this is called an essentially tau-equivalent measurement model), or when both the factor loadings and unique factor variances of the indicators of a given construct are restricted to be the same, which implies that the observed variables are fully exchangeable (this is called a parallel measurement model; see Traub, 1994).

For identification (i.e., to determine the unknown parameters of the measurement model uniquely), it is necessary to fix one loading per factor to one or, equivalently, to standardize the factor variances to one. When there are at least three indicators per factor, a congeneric factor

model is identified (even if there is only a single factor or if multiple factors are uncorrelated). When there are only two indicators per factor, a single-factor model is not identified unless additional restrictions are imposed, and multiple factors must be correlated for a two-indicators-per-factor model to be identified (at least some of the factor correlations have to be non-zero). When there is only a single indicator per factor, the associated unique factor variance cannot be freely estimated (i.e., has to be set to zero or another assumed value). Figure 4.1 shows an illustrative example of a congeneric measurement model with 8 observed measures and 2 correlated common factors in which the factor variances are standardized to 1 (so that all factor loadings can be free parameters).

Before a measurement model is analyzed in depth (as described below), it has to be ascertained that the specified model represents the available data reasonably well. The following two step-process is often employed. First, the overall goodness of fit of the specified model is examined using a chi-square test and various alternative fit indices. Often, the specified model is rejected based on the chi-square test, and rules of thumb associated with various alternative fit indices (RMSEA $\leq .08$, CFI $\geq .95$, TLI $\geq .95$, etc.) are employed to argue that the model is reasonable in a practical sense. Although it is true that the chi-square test is a rather stringent criterion of model fit, researchers are generally too quick to discount a significant chi-square statistic and should analyze models whose fit is judged to be questionable in greater detail. Second, if the fit of the model is deemed problematic, the model is re-specified in an effort to make it consistent with the data. This may involve dropping some indicators from the model (e.g., those that fail to load significantly on the target factor), specifying cross-loadings or correlated uniquenesses, combining factors, or introducing additional factors (including method factors). Model modification is usually based on an analysis of the residuals (discrepancies

between the observed variances/covariances and the variances/covariances implied by the specified model) and so-called modification indices, which estimate the improvement in the chi-square statistic when a fixed parameter is freely estimated or an equality constraint is relaxed.

A detailed measurement analysis consists of both an examination of reliability and convergent validity and an investigation of discriminant validity. With regard to the former, researchers generally want to know how well the indicators that were chosen to represent the construct of interest actually capture the construct. If very similar (exchangeable) items are used to measure a construct and the only source of error in measuring a construct is random measurement error, then the relationship between a construct and its indicators is called reliability. In contrast, when less similar items (e.g., indicators representing different methods for tapping the same construct) are employed and the unique factor associated with an item may contain sources of error other than random measurement error, the relationship between a construct and its indicators is called convergent validity. Usually, similarity is a matter of degree, not a difference in kind, so reliability and convergent validity are often used interchangeably.

Reliability and convergent validity can be assessed between individual indicators and a construct or between the entire set of indicators of a construct and the construct. Individual-item convergent validity can be assessed by the magnitude and significance of the factor loading of an indicator, but usually individual-item reliability (IIR) is reported, which is defined as the squared correlation between an item and the underlying construct (i.e., the proportion of the variance in an observed measure accounted for by the common factor). For the simple model in equation (1) it is given as

$$IIR_{x_i} = \frac{\lambda_{ij}^2 \varphi_{jj}}{\lambda_{ij}^2 \varphi_{jj} + \theta_{ii}} \tag{4.4}$$

A common rule of thumb is that IIR should be at least .5, although recommended values as low as .25 have appeared in the literature. When an item is related to multiple (correlated) factors, the formula is more complicated.

Researchers frequently report a summary measure of the individual-item reliabilities of the indicators of a given construct called average variance extracted (AVE; Fornell and Larcker, 1981). This is simply the average of the individual-item reliabilities of all the indicators of a construct. The usual rule of thumb is that AVE should be at least .5 (i.e., on average half of the variance of the indicators of a construct should be "substantive" or construct-related variance).

Researchers often report the reliability of an unweighted (or unit-weighted) composite of the indicators of a construct (i.e., the squared correlation between an unweighted sum, or average, of the indicators of a construct and the construct). This is called composite reliability (CR) and it can be computed as follows (assuming the measurement model is congeneric):

$$CR_{\sum x_i} = \frac{(\sum \lambda_{ij})^2 \varphi_{jj}}{(\sum \lambda_{ij})^2 \varphi_{jj} + \sum \theta_{ii}} \quad (4.5)$$

CR can be computed from the unstandardized solution, based on the variances and covariances of the observed measures, or the standardized solution in which the observed variables are transformed to have a variance of one; if the scales on which the observed variables are measured differ, the latter is preferable. CR is analogous to coefficient alpha, but based on somewhat weaker assumptions (i.e., the loadings do not have to be equal). Different guidelines for acceptable values of composite reliability and coefficient alpha are available in the literature (ranging from about .6 to .9). All rules of thumb are essentially arbitrary (citing somebody who proposed the rule doesn't make the rule less arbitrary), and it is best to take unreliability into account explicitly during the analysis, regardless of how reliable or unreliable the measures are. Of course, this requires that the estimated reliability accurately measure the "true" reliability.

With respect to discriminant validity, items should (primarily) measure the construct (or dimension of the construct) that they were meant to measure (discriminant validity at the item level), and there should also be discrimination at the construct level. Specifically, the sub-dimensions of a construct (if the construct was hypothesized to be multi-dimensional) should be distinct, and a construct (or its dimensions) should also differ from related constructs. As pointed out earlier, researchers usually relate each indicator to a single common factor (i.e., non-target loadings are restricted to zero *a priori*). In this case, discriminant validity at the item level is imposed *a priori*, but the tenability of this assumption can still be tested by looking at the modification indices associated with the loadings that are fixed at zero. If the modification index is (highly) significant and the expected parameter change (EPC), which estimates the change in the parameter if it were freely estimated, is non-negligible, then there is a problem with discriminant validity at the item level; often such indicators are dropped from the measurement model. At the construct level, researchers frequently report tests in which a model that restricts the correlation between two constructs to one is compared to a model in which the correlation is freely estimated. If the difference in the chi-square values of the two models is non-significant, the hypothesis that the correlation is equal to one cannot be rejected and there is a lack of discriminant validity at the construct level. This cumbersome procedure, besides being technically incorrect, is also unnecessary; the preferred approach is to construct a confidence interval around the factor correlations (the elements of the $\mathbf{\Phi}$ matrix are correlations when the factor variances are standardized to one) and check whether the confidence interval includes 1 (in which case discriminant validity is violated). The test of whether a correlation equals one may not provide strong evidence of discriminant validity because very high factor correlations

will be judged to differ from one when the confidence interval is narrow (e.g., due to a large sample size).

A generally stronger test of discriminant validity was suggested by Fornell and Larcker (1981). According to these authors, the squared correlation between two constructs should be smaller than the AVEs of the indicators measuring the two constructs. This criterion has some intuitive appeal (i.e., a construct should have more in common with its own indicators than with a presumably different construct), but it is not a statistical test (i.e., the AVE could be minimally higher in magnitude than the squared correlation), and by measuring constructs with nearly identical items, the test can be manipulated to favor the hypothesis of interest. It is possible and straightforward to statistically test whether the squared correlation between two constructs is lower than the AVE of the two sets of variables measuring the two constructs, but this is almost never done in practice.

4.1.2 Extensions of the basic reflective measurement model for continuous metric observed variables

The simple reflective measurement model, in which only the variances and covariances of the indicators and factors in a single sample are considered and the variation in each observed measure is assumed to be composed of two sources of variance (substantive variance due to a single common factor and unique factor variance, often equated with random error variance) can be extended in several ways (Baumgartner and Weijters, 2017). First, the assumption that each indicator is related to a single construct can be relaxed and more complex factor loading structures can be considered. In addition, the simple reflective measurement model assumes that the only source of covariation between items is shared substantive variance. However, often

there are many other sources of shared variance, particularly systematic measurement error (often called common method bias), which can induce covariation between the indicators (Podsakoff, MacKenzie, Lee, and Podsakoff, 2003). Second, the single-population measurement model can be extended to multiple populations, which also enables researchers to incorporate the means of indicators and factors. A very important application of multi-sample reflective measurement models with mean structures occurs in measurement invariance testing (Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000). The goal of measurement invariance testing is to ascertain whether measures are sufficiently similar in different populations (in terms of item intercepts and factor loadings) so that comparisons of substantive parameters (construct means, relationships between constructs) across populations can be conducted meaningfully. All these issues will be considered in this section.

4.1.2.1 More flexible loading patterns and unique factor covariance structures

In an independent cluster confirmatory factor analysis, the loadings of items on non-target substantive factors are restricted to zero *a priori*. This is a strong assumption which, if violated, may lead to poor model fit. In recent research, two approaches that relax this limiting assumption have been proposed. One is exploratory structural equation modeling (ESEM; Marsh *et al*., 2014), in which the usual congeneric measurement model is replaced with an exploratory factor analysis model (see also Baumgartner and Weijters, 2017). One way to specify such a model is to choose a reference indicator for each factor whose factor loading on the target factor is set to one and whose factor loadings on the non-target factors are set to zero; otherwise, all factor loadings are specified as free parameters. It is possible to statistically compare such an exploratory factor model with a confirmatory factor model, and it is likely that the exploratory

factor model will emerge as the preferred model, especially if the sample size, the number of factors, and the number of indicators per factor are relatively large. Although one could argue that the more complex exploratory factor model lacks parsimony, we recommend that researchers estimate such a model on their data and compare the resulting solution with the solution from a congeneric measurement model. The reason is that a misspecified congeneric factor model in which non-target loadings are incorrectly set to zero can have a substantial distorting effect on the findings (e.g., the factor correlations may be seriously inflated). In contrast, if the two solutions are similar, a researcher may want to retain the more parsimonious congeneric measurement model even when the fit is worse based on statistical criteria (especially when the non-target loadings, although significant, are small in magnitude).

Muthén and Asparouhov (2012) proposed a second approach to modeling a more flexible factor pattern for the substantive factors using Bayesian Structural Equation Modeling (BSEM). In this approach, all factor loadings (both target and non-target loadings) are specified as free parameters, but the model is identified by using informative priors with a small variance for the non-target loadings (e.g., a normal prior with a mean of zero and a variance of .01 for the standardized loadings, which implies a 95 percent confidence interval for the loadings ranging from -.2 to +.2). Although little experience with this method is available to date, it is a novel and promising approach to specifying measurement models.

So far in the discussion the focus has been on whether items should be allowed to load on multiple substantive factors. Often, there are strong reasons to suspect that a single substantive factor or multiple correlated substantive factors are not the only source of covariation between indicators (of either the same construct or different constructs). In particular, there is now a sizable literature on common method bias, which strongly suggests that various factors unrelated

to the substantive constructs of interest can induce a correlation between the indicators (e.g., Podsakoff *et al*., 2003). In general, characteristics of the respondent, properties of the items, and features of the survey instrument and the survey context may lead to shared method variance in items. Characteristics of the respondent include response styles (esp. acquiescence, extreme, and midpoint response style) and response sets (esp. social desirability), which were discussed earlier, but there are other individual difference variables as well (e.g., need for consistency). Properties of the items consist of attributes of the question (e.g., the keying direction of the question) and attributes of the response scale (e.g., the use of the same response scale in items measuring the same or a different construct). Finally, features of the survey instrument and survey context include the positioning of items in the questionnaire and the mode of data collection (e.g., telephone interview, online survey, etc.).

Many different models that incorporate method effects have been suggested, which differ in terms of (a) whether method effects are measured explicitly or modeled implicitly; (b) whether method effects are considered at the factor level or at the level of individual items; and (c) whether method effects are modeled via method factors or correlated uniquenesses. Figure 4.2 presents four prototypical method effect models (MEMs) that can be derived using a decision tree based on these criteria. Figure 4.3 shows illustrative examples of these models assuming either that an explicit measure of acquiescent response style (ARS), as one possible cause of method effects, is available, or that ARS is modeled implicitly based on respondents' answers to both regular and reversed versions of items measuring the same substantive construct. For concreteness, the illustrative example posits two substantive factors, as well as two regular and two reversed items per substantive factor, but other types of factor models are of course possible.

In MEM-1 and MEM-2, a direct measure of the assumed method effect is available. This generally requires that the method effect was hypothesized *a priori* and that items measuring the method effect in question were included in the questionnaire. Ideally, the items measuring the method effect should be completely unrelated to the substantive construct of interest so that substance and method are unconfounded. If multiple method effects are hypothesized, multiple measures of method effects can be included in the model. In the illustrative example of Figure 4.3, the method effect is due to ARS, and a suitable measure of ARS can be derived from the incidence or strength of agreement with a set of items that are heterogeneous in content (so that agreement responses truly measure style and not content, since the items lack common content). If there is measurement error in the overall method effect measure, unreliability can be accounted for in the model, although this is not shown in the example of Figure 4.3. MEM-1 and MEM-2 differ in that with the former the explicit method-effect measure is related to the substantive factors, whereas in the latter the explicit method-effect measure is related to the individual items measuring the substantive factors. In MEM-1 the paths from the method effect measure simply show whether the substantive factors are contaminated by method variance, whereas in MEM-2 method variance is removed from the individual items and the purged items are related to the substantive factors they are supposed to measure. In this way, method variance cannot contribute to the common substantive variance shared by the items. In general, it is preferable to take into account method variance at the level of individual items.

In MEM-3 and MEM-4, method effects are modeled implicitly. In general, it is dangerous to infer method effects from the substantive items themselves, because substantive and method variance cannot be distinguished clearly and will normally be confounded. However, in special cases implicit method effects can be modeled. One example is a situation in which both regular

and reversed items measuring the same construct are available. For example, if somebody agrees with the item “I am satisfied with this brand” and also agrees with the item “I am dissatisfied with this brand”, this person’s responses are presumably not based on the substantive content of the items. A more plausible hypothesis might be that this pattern of responding reflects a tendency to agree with items regardless of content (although there could be other reasons).

MEM-3 and MEM-4 both model method effects at the individual-item level. However, in MEM-3 method effects are modeled with method factors, whereas in MEM-4 method effects are modeled with correlated uniquenesses. Both specifications have strengths and weaknesses (see Lance, Noble, and Scullen, 2002), but in many circumstances MEM-3 is the preferred specification. In MEM-3, the two method factors are specified to be correlated, but one can also test whether the two method factors are uncorrelated or perfectly correlated, which implies that there is a single acquiescence factor that underlies people’s responses to items measuring different constructs. While MEM-3 assumes that acquiescence for one construct is correlated with acquiescence on another construct (presumably, the correlation should be positive), MEM-4 implies uncorrelated method effects, which may not be very realistic in the present context.

There is another specification of method effects, similar to MEM-3, which is available when each of several constructs is measured by each of several methods. This model has figured prominently in the construct validation literature, is known as the multitrait-multimethod (MTMM) approach, and was originally introduced to simultaneously investigate convergent and discriminant validity (Campbell and Fiske, 1959). An updated version based on confirmatory factor analysis will be briefly described here (see Bagozzi and Phillips, 1991, for details). To make the discussion more concrete, consider the case where items varying in their keying direction (regular or positively worded items vs. reversed or negatively worded items) serve as

the two methods, and an equal number of both types of items (two in the present case) is used to measure two constructs. Although items varying in their keying direction hardly satisfy the requirement that the methods used should be maximally different (Campbell and Fiske, 1959), we employ these "methods" in order to compare and contrast the approach with the method models described earlier. As shown in Figure 4.4, the items sharing the same keying direction load on the same method factor (M1 and M2, respectively) and the two method factors are allowed to correlate freely, but the method factors are specified to be uncorrelated with the two (freely correlated) substantive factors (A and B). In contrast to MEM-3, the method factors do not model acquiescence directly. Instead, the method factors represent shared variance due to the use of the same method of measurement (i.e., either positive or negative wording of the items). However, if the two method factors are positively correlated, which implies that respondents who (dis)agree with the positively worded items also (dis)agree with the negatively worded items (assuming reversed items were not recoded), this would be consistent with the hypothesis that acquiescence is a source of the commonality between the two method factors.

Widaman (1985) proposed a taxonomy of structural models for MTMM data which allows researchers to conduct the following model comparisons (using chi-square difference tests): (a) an omnibus test of convergent validity, which compares the model including both freely correlated substantive and freely correlated method factors (the so-called correlated trait-correlated method or CTCM model) with a model including only method factors; (b) an omnibus tests of discriminant validity, which compares the CTCM model with a model specifying perfectly correlated substantive factors (or a single substantive factor) and freely correlated method factors; (c) an omnibus test of the discriminability of methods, which compares the CTCM model with a model specifying freely correlated substantive factors and perfectly

correlated method factors (or a single method factor); and (d) an omnibus test of the presence of method effects, which compares the CTCM model with a model including only substantive factors. Often, these model comparisons will show that the CTCM model will have the best fit, which implies that, in an overall sense, there is evidence of convergent and discriminant validity, that the methods are discriminable, and that method effects are present. Follow-up tests can then be conducted to determine the proportion of substantive, method, and error variance in each indicator (i.e., the proportion of the total variance in an item accounted for by the substantive factor, the method factor, and unique sources of variance), and the discriminant validity of the constructs can be checked using the methods described earlier (see Bagozzi and Phillips, 1991, for details).

The CTCM model in Figure 4.4 can be modified in various ways. First, method factors can be specified to be freely correlated or perfectly correlated, as already mentioned, but they could also be uncorrelated. Second, method effects may only be present for the regular items or for the reversed items, in which case only one method factor would have to be included. Third, MTMM models often suffer from convergence problems and improper solutions, which has led to efforts to propose alternative specifications. One such specification is the model of correlated uniquenesses, in which the unique factors of items that share the same keying direction are allowed to correlate. As in the case of method factor models, correlated uniquenesses can be specified for the regular items, for the reversed items, or both. Another specification is a variant of the CTCM model proposed by Eid (2000) in which one method is chosen as a comparison standard so that the number of method factors is one fewer than the number of methods. The resulting model is called the correlated trait-correlated methods minus one or CTC(M-1) model.

Research shows that the CTC(M-1) model is identified under more general conditions than the usual CTCM model for MTMM data and overcomes several of its other limitations.

It should be noted that the factors that aim to capture method variance in the CTCM model often seem to also capture substantive variance to some extent, making it hard to unequivocally interpret the "method" factors in the model. For this reason, the previously discussed model using an ARS factor with same-sign loadings for reversed and nonreversed items may often be preferable.

4.1.2.2 Multi-sample reflective measurement models for continuous metric observed variables

The reflective measurement model specified in equation (4.2) applies to a single population. Sometimes, researchers want to compare measurement models across several populations (e.g., males vs. females or respondents from different countries). Even if a researcher is not primarily interested in comparing measurement models across several populations, evidence of measurement invariance has to be provided whenever the magnitude of means or the strength of relationships between constructs is to be compared across several populations, otherwise the comparisons conducted may be misleading or meaningless.

Apart from enabling comparisons of measurement models across several populations, a multi-sample measurement model also allows researchers to incorporate the means of observed variables and factors. Thus, the multi-sample measurement model with a mean structure can be stated as follows:

$$\boldsymbol{x}^g = \boldsymbol{\tau}^g + \boldsymbol{\Lambda}^g \boldsymbol{\xi}^g + \boldsymbol{\delta}^g \tag{4.6}$$

where $\boldsymbol{\tau}$ is an I x 1 vector of equation intercepts and the other terms were defined earlier. The superscript g (g for group) signals that the measurement model applies to one of G populations.

Under appropriate assumptions (see the earlier discussion), the corresponding mean and covariance structures are:

$$\boldsymbol{\mu}^g = \boldsymbol{\tau}^g + \boldsymbol{\Lambda}^g \boldsymbol{\kappa}^g \quad (4.7)$$

$$\boldsymbol{\Sigma}^g = \boldsymbol{\Lambda}^g \boldsymbol{\Phi}^g \boldsymbol{\Lambda}'^g + \boldsymbol{\Theta}^g \quad (4.8)$$

where $\boldsymbol{\mu}$ is the expected value of $\mathbf{x}$ and $\boldsymbol{\kappa}$ is the expected value of $\boldsymbol{\xi}$ (i.e., the vector of latent means of the common factors or constructs). To identify the covariance structure one loading per factor should be set to one; in contrast to single-group models, the factor variances should not be standardized because it would impose the unrealistic assumption that the factor variances are equal across groups in models of metric and scalar invariance (see below). The means part can be identified in different ways, but one possibility is to set the intercept of the reference indicator or marker item (the item whose loading is fixed at one) to zero.

The model in equations (4.7) and (4.8) consists of five parameter matrices, three of which contain measurement parameters ($\boldsymbol{\tau}^g$, $\boldsymbol{\Lambda}^g$, $\boldsymbol{\Theta}^g$) and two of which contain substantive parameters ($\boldsymbol{\kappa}^g$, $\boldsymbol{\Phi}^g$). In practice, researchers who want to compare relationships between constructs across multiple populations will be primarily interested in comparing path models across groups, but directed paths are completely determined by the variances and covariances in $\boldsymbol{\Phi}$. In order to establish measurement equivalence, the following models should be compared (Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000). The most basic requirement of invariant measurement is that the factor structure is the same in the different populations (i.e., the number of observed variables and factors, and the pattern of factor loadings that are restricted to zero, is the same across groups). This is called configural invariance. If structural relationships between constructs (e.g., a directed path from construct A to construct B) are to be compared across groups, metric invariance (i.e., equality of factor loadings) has to be satisfied. Finally, if factor

means are to be compared across groups, scalar invariance (equality of item intercept, in addition to equality of factor loadings) has to be satisfied as well. To test whether measurement equivalence of a given type holds, chi-square difference tests can be conducted. For metric invariance the model in which the factor loadings are equated across groups is compared with the configural model, and for scalar invariance, the model in which the item intercepts are equated is compared with the metric invariance model. If the chi-square difference test (i.e., the difference in chi-square values between two models relative to the difference in degrees of freedom) is significant, metric or scalar invariance is rejected. When this is the case, the parameter restrictions that are unjustified have to be relaxed; this is usually done with the help of modification indices.

Since full metric and full scalar invariance (i.e., equality of all parameters in τ and Λ) are often violated (especially the latter), weaker conditions of measurement invariance are desirable. Steenkamp and Baumgartner (1998) show that, at a minimum, two items per factor should satisfy metric invariance (for comparisons of structural relationships) and scalar invariance (for comparisons of factor means); this is called partial measurement invariance. Of course, it is preferable if most of the indicators of a given factor exhibit measurement equivalence.

Exact metric and scalar invariance of (even a subset of) items is sometimes difficult to achieve in practice. This has encouraged the development of methods that implement approximate measurement invariance. In addition, if the number of groups to be compared is large, considering groups as a fixed mode of variation is cumbersome, and methods have been developed that treat groups as a random mode of variation. Details about these approaches are provided in Muthén and Asparouhov (2018). It should also be noted that when the factor models to be compared are complex and the number of groups is large, strict null hypothesis significance

testing becomes less useful and a modeling perspective may be preferable. This means that the focus should be on fitting models that are sufficiently invariant across groups based on practical fit indices such as RMSEA or BIC (Weijters, Puntoni, and Baumgartner 2017).

4.1.3 The reflective measurement model for discrete ordinal observed variables

In practice observed variables are usually not continuous (slider scales are one exception), but if a variable can assume many discrete values, little is lost by treating it as continuous. However, usually, the number of discrete categories is quite small. Often, the data are discrete by design, because the number of response options is limited (e.g., when a 5- or 7-point Likert agree-disagree scale is used). Even when the number of potential response options is larger (e.g., when a researcher uses a 100-point scale), observed responses often cluster around a limited number of values (e.g., round numbers). In addition to the issue that data are often discrete rather than continuous, many response scales used in marketing are not interval scales. For example, the difference between agree and strongly agree is probably not the same as the difference between disagree and neither agree nor disagree. Theoretically, 5- or 7-point agree-disagree scales should be treated as discrete and ordinal (also called ordered-categorical), but from a practical perspective the question is whether treating such scales as continuous and metric has serious shortcomings. Below, we will first discuss approaches that model observed variables as ordered-categorical and then consider the issue of when ordered-categorical data can be treated as continuous for practical purposes, including what adjustments need to be made in order for the statistical tests to be more accurate. Finally, we will also briefly discuss the issue of measurement invariance testing when the data are ordered-categorical.

4.1.3.1 Item response theory (IRT) and related approaches to modeling ordered-categorical data

One way in which the discreteness and ordinal nature of the data can be modeled explicitly is by assuming that the observed variables are discretized versions of underlying continuous variables. For example, the observed variables may be 5-point strongly disagree to strongly agree responses, but they are imperfect measures of a respondent's strength of (dis)agreement with an item. The factor model is similar to equation (4.1), but the previously observed x_i is now a latent response tendency x_i^*:

$$x_i^* = \lambda_{ij}\xi_j + \delta_i \tag{4.9}$$

The observed variable x_i is a discretized version of x_i^* such that $x_i = k$ if $\nu_{k-1} < x^* \leq \nu_k$, where K is the number of response options (k = 1, 2, …, K) and ν_1 to ν_{K-1} are thresholds to be estimated, where $-\infty = \nu_0 < \nu_1 < \cdots < \nu_{K-1} < \nu_K = \infty$. It can be shown that this model is equivalent to the graded response model of item response theory or IRT (Samejima, 1969), in which the choice of a response in one of the K categories is modeled by (K-1) sigmoid curves (called operating characteristic curves or cumulative response curves) that express the probability of providing a response of k or higher, that is,

$$P(x_i \geq \mathrm{k}|\xi_j) = F(\alpha_{ij}\xi_j + \beta_{ik}) = F\left(\alpha_{ij}(\xi_j - \gamma_{ik})\right) \tag{4.10}$$

where k = 2, …, K since $P(x_i \geq 1) = 1$. In equation (4.10), F is either the normal or logistic cumulative distribution function, the α_{ij} are slope or discrimination parameters, and the β_{ik} and γ_{ik} are item-specific intercept or threshold parameters for the various response categories k. Since the α_{ij} are constant for the different response categories, the slopes of the different sigmoid curves are parallel. The probability of a response in the k^{th} interval is given by the difference of $P(x_i \geq k)$ and $P(x_i \geq k+1)$, that is,

$$P(x_i = \mathrm{k}|\xi_j) = P(x_i \geq \mathrm{k}|\xi_j) - P(x_i \geq \mathrm{k}+1|\xi_j) \quad (4.11)$$

The curves describing the relationship between $P(x_i = k|\xi_j)$ and ξ_j are called category characteristic, category response or item characteristic curves.

To identify the model, it is necessary to choose a scale for x_i^* and ξ_j. Different parameterizations can be employed, but one possibility is to (a) set the variance of δ_i to unity (called the conditional parameterization of the continuous response variable by Kamata and Bauer, 2008 or the theta parameterization in Mplus) and (b) constrain the variance of ξ_j to one (called the standardized parameterization of the latent construct by Kamata and Bauer, 2008). Instead of constraining the variance of δ_i to unity, one can also set the variance of x_i^* to one, which is called the delta parameterization in Mplus.

Muraki (1990) developed a modified graded response model specifically designed for Likert-type items in which there is a separate γ_i parameter for each item, but the distances between adjacent thresholds are the same across items (i.e., $\gamma_{ik} = \gamma_i + c_k$). A special case of the ordinal factor or graded response model is the binary factor or two-parameter binary IRT model. This model is used when there are only two response options (e.g., yes vs. no, or agree vs. disagree).

In general, the IRT approach considers a respondent's entire response pattern to all items measuring a given construct, so it is a full information categorical variable procedure. Although in theory full information methods have certain advantages, in practice limited information procedures, which use only some of the information contained in the raw response data, do as well or even better than full information approaches (Rhemtulla, Brosseau-Liard, and Savalei, 2012). Limited information categorical estimation methods are generally based on the following two steps. First, the correlations between the continuous variables x_i^* that are posited to underlie

the discrete observed variables x_i are recovered using the univariate and bivariate frequencies of the x_i. These correlations are called polychoric correlations (or tetrachoric correlation if the variables are binary). Second, the polychoric correlations (rather than the variances and covariances of the observed variables) are used as inputs to a conventional confirmatory factor analysis. Although several limited information methods are available, depending on the weight matrix used during estimation, Rhemtulla *et al*. (2012) recommend categorical (unweighted) least squares estimation with robust correction of test statistics and standard errors (esp. when the sample size is medium to small).

4.1.3.2 When can continuous methods be used with ordered-categorical data?

An alternative to full or limited information categorical variable methods is the use of continuous normal theory maximum likelihood procedures coupled with robust correction of test statistics and standard errors for violations of normality (because categorical data are by definition non-normal). Since the discreteness and ordinal nature of the input data is not modeled explicitly (the resulting non-normality is only considered in a general way), the model is misspecified and the parameter estimates will be biased, but once the number of response categories is sufficiently high and the continuity assumption becomes more plausible, it is hoped that the resulting parameter estimates (esp. when coupled with robustness corrections) will approximate the results obtained with more appropriate categorical variable methods.

Rhemtulla *et al*. (2012) presented an excellent review of prior research on the performance of continuous and categorical estimation procedures in the presence of discrete ordinal data, and they conducted their own extensive simulation of factors influencing the estimation results. Specifically, they compared limited information categorical least squares with

continuous normal theory maximum likelihood estimation (both with robust corrections), and they studied the following influences on the quality of estimation: (a) model size (a two-factor CFA model with 5 or 10 indicators per factor); (b) number of categories (2, 3, 4, 5, 6, or 7); (c) sample size (100, 150, 350, 600); (d) threshold symmetry (symmetric, moderately asymmetric, and extremely asymmetric); and (e) normality of the distribution of the variables underlying the discrete variables (normal, nonnormal). The outcome variables investigated included the incidence of convergence failures, bias and efficiency of parameter estimates, bias and coverage of robust standard errors, and type I error and power of (robust) test statistics of overall model fit. Their findings showed that, with two to four response categories, continuous methods can be problematic. However, with five to seven response categories, the conclusion was that "reliance on continuous methodology in the presence of ordinal data will produce acceptable results" (p. 371). In marketing it is rare to find empirical studies in which scales with fewer than five response options are used (unless the data are binary, in which case methods designed for binary data should be used), so it seems safe for researchers to use continuous methods with robust corrections to test statistics and standard errors, especially when the distributions of the variables are approximately symmetric.

4.1.3.3 Multi-sample reflective measurement models for ordered-categorical variables

If researchers intend to conduct comparisons of means or structural relationships across multiple populations, measurement invariance is just as important for ordered-categorical data as for continuous data. However, the issues are much more complex, and at present it is not entirely clear how measurement equivalence should be ascertained and, more importantly, what type and

degree of measurement invariance is necessary in order for comparisons of means and structural relationships across groups to be meaningful.

Millsap and Yun-Tein (2004) were one of the first to consider the assessment of factorial invariance for ordered-categorical observed variables, and they suggested that one way to identify the configural invariance model, which serves as the baseline model for further model comparisons, is to impose the following identification conditions (assuming a congeneric measurement model): (a) all item intercepts in all groups have to be set to zero (since thresholds and intercepts cannot be simultaneously free model parameters); (b) a marker (or reference) item with a loading of 1 on the target factor has to be chosen for each latent factor and group; (c) the variances of the x_i^* have to be set to one and the factor means have to be set to zero in one of the groups (since the item intercepts and factor means are zero, this implies that the means of the x_i^* are zero in that group); and (d) one threshold per item and a second threshold for the marker item of each factor has to be invariant across all groups. The resulting model serves as the baseline model, and if this baseline model fits the data, the invariance of loadings and thresholds can be tested.

Unfortunately, as pointed out in a recent article by Wu and Estabrook (2016), different ways of identifying the baseline model (the identification conditions suggested by Millsap and Yun-Tein, 2004, are just one possibility) lead to different scales for the x_i^*, which affects subsequent model comparisons in which additional invariance constraints are imposed. Wu and Estabrook (2016) discuss different identification conditions for various combinations of threshold, loading, intercept, and unique variance invariance, but it is unlikely that full invariance of thresholds, loadings, and intercepts will hold in practical applications and little is

known about the extent to which the invariance constraints of a given type can be relaxed for desired comparisons of parameters to remain meaningful.

4.1.4 An empirical illustration of measurement analysis for reflective indicator models

Since some of the ideas related to reflective measurement models may not be familiar to all readers, we will present an illustrative example dealing with the measurement of consumers' material values (the data and Mplus command files are available at https://github.com/HansBaum129/MarketingMeasurement and a summary of the analyses is provided in Table 4.2). Richins and Dawson (1992) proposed a conceptualization of materialism consisting of three dimensions – possession-defined success, acquisition centrality, and acquisition as the pursuit of happiness (referred to as the success, centrality, and happiness dimensions below) – and they developed an 18-item scale to assess individual differences in materialism (with 6, 7, and 5 items per dimension, respectively). Although the scale contains 8 reversed items, it is not balanced by dimension. For the following analyses, we modified the original scale slightly by (1) dropping one of the reversed centrality items (which was also dropped in Richins', 2004, revision of the original scale), (2) modifying one of the success items to make it a reversed item (from 'The things I own say a lot about how well I'm doing in life' to 'I don't think that the things I own say a lot about how well I'm doing in life'), and (3) adding a new reversed happiness item (since there were only two reversed happiness items in the original scale, i.e., 'I am happy with the things I already own; I don't need additional luxuries'). The resulting scale is balanced by dimension, with an equal number of regular and reversed items per dimension. Data are available for 554 respondents (undergraduate students), who completed a survey administered via Qualtrics (which contained a variety of other questions) for course

credit. The order of the 18 items was randomized for each respondent, and participants provided their responses on 5-point agree-disagree scales (1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree).

We first conducted exploratory maximum likelihood factor analyses specifying 1 to 6 oblique factors, using Promax for factor rotation and treating the data as continuous and metric. The 6-factor model achieved an acceptable fit with a non-significant p-value (χ^2 (60) = 77.55, p = .063), but the factor pattern matrix did not have a straightforward interpretation. The fit of the 3-factor solution was relatively poor (χ^2 (102) = 239.99, p < .0001). Although the happiness items tended to load on the same factor (esp. the regular items), only the regular centrality items loaded on the second factor and only the reversed success items loaded on the third factor. Similar results were obtained when the observed variables were treated as ordered-categorical.

Researchers would normally estimate a confirmatory three-factor model, given that the scale was designed to measure three dimensions of materialistic values. Since the exploratory factor model with three factors did not fit well, a more restrictive confirmatory factor model in which each item is allowed to load only on its target factor will likely have an even poorer fit. This was indeed the case: χ^2 (132) = 415.24 (p < .0001), RMSEA = .062 (90% CI .056 to .069), SRMR = .052, CFI = .876, TLI = .856. All factor loadings were highly significant, although only 13 of 18 standardized loadings exceeded .5. Individual-item reliabilities ranged from .025 to .644, average variance extracted for the indicators of success, centrality, and happiness was .290, .252, and .372, respectively, and the corresponding composite reliability estimates were .705, .646, and .770. The factor correlations were .818 (between success and centrality), .720 (between success and happiness), and .629 (between centrality and happiness). Although the factor correlations differed from one (i.e., none of the confidence intervals around the factor

correlations included one), the Fornell and Larcker criterion suggested a lack of discriminant validity (since the squared factor correlations exceeded the average variance extracted estimates).

Richins and Dawson (1992) did not provide a lot of information about their (confirmatory) scale validation efforts, but they mentioned that, across three data sets, the coefficient alpha estimates ranged from .74 to .78 for the six success items; .71 to .75 for the seven centrality items; and .73 to .83 for the five happiness items. Factor correlations ranged from .39 to .79, and the adjusted goodness-of-fit indices ranged from .86 to .88. The composite reliability estimates are somewhat lower in our data (the corresponding coefficient alpha estimates were .70, .63, and .76 for success, centrality, and happiness, respectively); the highest factor correlation in their data is comparable to the highest factor correlation in our data; and the adjusted goodness-of-fit index (which is no longer a recommended fit index) is also similar in our data (.885). Richins (2004) summarizes re-analyses of 15 different data sets and reports mean RMSEA, CFI, and TLI estimates of .07, .86, and .84, and mean alphas of .77, .72, and .78 for success, centrality, and happiness. The present results are similar to these findings.

Some researchers might be tempted to conclude that the fit of the three-factor model is adequate, based on fit indices such as RMSEA, and that the composite reliabilities indicate reasonable internal consistency (with the exception of centrality). However, it is of interest to investigate why the highly restrictive three-factor model (in which the observed variables load only on the factor that they were designed to measure, and the hypothesized three substantive factors are the only source of covariation between the items) is inconsistent with the data.

To examine the assumption of zero non-target loadings, one can look at the modification indices of the loadings that are constrained to zero. A total of 10 non-target loadings (out of 36) had significant modification indices, and 4 were substantial (greater than 10). If 11 (originally

zero) non-target loadings with significant modification indices are sequentially freed, the overall chi-square value (with 121 degrees of freedom) is reduced to 290.76. Although the revised model has a better fit (i.e., restricting all non-target loadings to zero contributes to the observed lack of fit), data-based model modification is an a-theoretical approach that does not necessarily lead to a preferred specification that will hold up in future studies (MacCallum, 1986; MacCallum, Roznowski, and Necowitz, 1992).

Bayesian Structural Equation Modeling (BSEM) is another approach for identifying salient non-target loadings. We estimated a model in which all factor loadings were free parameters, but the non-target loadings were identified by using informative normal priors with a small variance of .01, which implies a 95 percent confidence interval for the standardized loadings ranging from -.2 to +.2. Only three non-target loadings had confidence intervals that did not include zero, and these were also the non-target loadings that were freed first in the sequential model respecification process based on modification indices.

Prior research has shown that reversed items often share method variance (Weijters, Baumgartner, and Schillewaert, 2013). In other words, items may covary because they measure the same substantive factor (success, centrality, and happiness in the present context), but also because they share the same keying direction (i.e., they are either regular or reversed items). Method effects defined by a common keying direction can be modeled using either method factors or correlated uniquenesses, and method factors or correlated uniquenesses can be specified for the regular items, for the reversed items, or both. Furthermore, if separate method factors are used for the regular and reversed items, they can be specified to be uncorrelated, freely correlated, or perfectly correlated. If the two method factors are perfectly correlated, the resulting model is equivalent to a model with a single general method factor. The model with a

single method factor is formally identical to a so-called bi-factor model in which a general (usually substantive) factor underlies all items, but there are also subfactors for subsets of items. When half of the items are reversed and reversed items have not been recoded to establish a uniform keying direction across all items, regular items should have a positive loading on the underlying substantive factor and reversed items should have a negative loading. If both regular and (non-recoded) reversed items have positive loadings on the general factor, the general factor cannot be a substantive factor (because respondents do not discriminate between regular and reversed items and indicate either agreement or disagreement with items regardless of keying direction). If there are multiple substantive factors and multiple method factors, the model is a multitrait-multimethod model. Finally, the loadings on method factors can be freely estimated, or the loadings can be restricted to be the same across items loading on the same method factor. When there is a single method factor and the loadings are specified to be the same, the resulting model has been called the random intercept model (Maydeu-Olivares and Coffman, 2006).

In principle, separate method factors could be considered for each dimension of the construct. However, when the items refer to the same construct (even though they measure different dimensions of the construct) and appear in close proximity in a questionnaire (e.g., on the same screen in online administration), it is unlikely that separate method effects are needed to properly account for method variance in the items, and such models will not be considered here.

Method effects (due to keying direction in the present case) can also be modeled with correlated uniquenesses. This means that the unique factors (“errors”) of items that share the same keying direction are allowed to correlate. Correlated uniquenesses can be specified for the regular items, the reversed items, or both. In contrast to models with two method factors, which could be allowed to correlate, separate correlated uniquenesses for regular and reversed items

cannot be correlated (i.e., the model is similar to a model with uncorrelated method factors). However, models with correlated uniquenesses impose no assumption of unidimensionality of method effects. If the correlated uniquenesses are restricted to be the same for all items sharing the same keying direction, the model is identical to the model with two uncorrelated method factors whose loadings are specified to be the same for a given method factor.

We estimated the different method effect models discussed above, and relevant fit information is reported in Table 4.2. In general, the models with correlated uniquenesses fit the data better than the models without method effects (except when fit assessment is based on BIC), but they require the estimation of many parameters and thus lack parsimony (i.e., these models have far fewer degrees of freedom). Based on the fit indices that take into account model parsimony (RMSEA, TLI, and BIC), they perform more poorly than some of the models with method factors. In the model with correlated uniquenesses for both regular and reversed items, 64 of 72 correlated uniquenesses are positive, which suggests that the correlated uniquenesses are not simply garbage parameters but capture a shared source of covariation between regular and reversed items, respectively. It is interesting to note that correlated uniquenesses among the regular items are more important in accounting for covariation among the items compared to correlated uniquenesses among the reversed items (parenthetically, only one correlated uniqueness has a negative estimate for the regular items). Still, models with correlated uniquenesses do not seem to provide an appealing (i.e., parsimonious) representation of the data in the present context.

All method factor models fit substantially better than the models without method effects. The models in which the method loadings are restricted to be the same are more parsimonious than the models in which the method loadings are freely estimated, and based on the fit indices

that impose a penalty for lack of parsimony, the fit is similar to, and sometimes even better than, the fit of the models in which all method loadings are freely estimated. The best-fitting model (based on RMSEA and TLI, though not BIC) has two correlated method factors and free method loadings, but it is followed closely by the model with two uncorrelated method factors and free loadings. The model with one method factor and equal loadings also has a relatively good fit and is the model with the lowest BIC. The substantive implications of these models differ substantially, however. In the model with freely correlated "method" factors, the correlation is negative (-.586), which suggests that the two factors actually capture materialistic values in an overall sense, independently of the three specific dimensions of materialistic values, although there is apparently something particular to items sharing the same keying directions. In contrast, in the model with one method factor and equal loadings, the method factor clearly captures (dis)agreement with the materialistic values items regardless of keying direction. The model with two uncorrelated method factors represents a middle ground: the two method factors capture communalities within regular and reversed items, but the substantive interpretation is not clear.

Overall, when evaluating model fit in combination with interpretability, the model with three trait factors and one method factor with equal loadings seems to be the preferred model (since the model with two "method" factors cannot clearly disentangle content and method variance). It is interesting to note that the ambiguity in the interpretation of correlated method factors was pointed out as early as 1989. Specifically, Marsh (1989, p. 357) suggested that "the implicit assumption that so-called method factors represent primarily the effects of method variance" is "often implausible" and that so-called method factors may represent "trait variance in addition to or instead of method variance." Furthermore, Marsh (1989) speculated that this

problem may be especially likely in models with correlated method factors. The present results confirm Marsh's observations.

To get further insights into the presence of method effects, we also estimated some of the method effects models depicted in Figure 4.3. To investigate the effect of directly measured method effects on respondents' ratings, we investigated the influence of acquiescent responding and impression management on people's endorsement of materialistic values. The acquiescence response style (ARS) measure was computed from 16 items that are free of common content, which should yield a "pure" measure of acquiescence (style). Specifically, for each of the 16 items a response of 5 (or strongly agree, regardless of the keying direction of the item) was scored as 2 and a response of 4 (or agree) was scored as 1, and the individual-item ARS scores were then averaged across the 16 items. To measure impression management (IM, or social desirability) we used 10 items from Paulhus' (1991) BIDR scale (the items are shown in Steenkamp *et al*., 2010). Respondents rated each item on 5-point strongly disagree to strongly agree scales, and the IM score is computed as the average response to the 10 items (after reverse-coding the reverse-worded items).

We specified a three-factor congeneric measurement model for the 18 MVS items (without recoding the reversed items) and treated both ARS and IM as antecedents of the 18 individual items. As discussed earlier, it is preferable to take into account method effects at the individual-item level (i.e., MEM-2 is preferable to MEM-1). The fit of this model was similar to the fit of the model without method effects: χ^2 (132) = 393.46 ($p < .0001$), RMSEA = .060 (90% CI .053 to .067), SRMR = .045, CFI = .892, TLI = .845. Tests of the null hypothesis that the effects of ARS or the effects of IM on item scores were simultaneously equal to zero indicated that both null hypotheses could be rejected at high levels of confidence. Since reversed items

were not recoded, a positive effect of ARS on responses to individual items provides evidence of acquiescent responding. Out of 18 effects, 14 were positive, and 7 were significant using a two-sided test at $\alpha = .05$; only one of the 4 negative effects was significant. For IM, a positive effect of IM on the regular items and a negative effect of IM on the reversed items provides evidence of impression management. All effects but one had the right sign (one effect was 0), and 16 were significant. These findings indicate that people's responses to the MVS items are influenced to some extent by acquiescent response tendencies and impression management.

We also estimated MEM-3 (see Figure 4.3), initially specifying 3 correlated implicit ARS factors (one for each dimension of MVS). However, the correlation between the first two ARS factors (for success and centrality) was 1, so we respecified the model and considered only two method factors. The fit of this model, either with freely estimated loadings or with equal loadings (see Table 4.2), was comparable to, but slightly better than, the model with one method factor (the correlation between the two method factors in the model with free method loadings was .651, with a confidence interval ranging from .437 to .865). All method loadings except one were positive in the model with free method loadings (although only 6 were significantly positive), and in the model with equal factor loadings the loadings for both factors were significant.

To provide further insights into the implicitly modeled ARS factors, we included ARS and IM in the model with two correlated method factors (MEM-3). Specifically, we regressed the three MVS dimensions as well as the two implicit ARS factors on both the direct ARS measure and IM. Success, centrality and happiness were unrelated to measured ARS, but significantly positively correlated with IM. The two implicit method factors were significantly correlated with measured ARS, but unrelated to IM. These results confirm that the implicit method factors indeed represent ARS (because the inferred ARS factors are significantly correlated with a direct

measure of ARS based on completely different content-free items), and the findings additionally indicate that self-ratings on materialism are influenced by impression management.

If reliability is recalculated from the model in which ARS and IM are controlled at the item level, composite reliability (average variance extracted) is .70, .64, and .77 (.28, .25, and .37) for the success, centrality, and happiness dimensions, respectively. The factor correlations are .81 between success and centrality, .70 between success and happiness, and .60 between centrality and happiness. Both the reliabilities and the factor correlations decrease somewhat when ARS and IM are controlled at the item level, but the differences are small. A major advantage of using reversed items is that they control for acquiescence response tendencies. Although acquiescence did influence response to individual items, the use of a balanced scale in which half of the items were regular and half were reversed eliminates the distorting influence of acquiescence on factor scores.

The analyses in Table 4.2 are based on maximum likelihood estimation, which assumes that the data are continuous and metric. If a robust correction to the fit statistics and standard errors is used, the overall model fit improves somewhat. If categorical (unweighted) least squares estimation with robust correction of test statistics and standard errors is used, the overall model fit deteriorates, but the reliability of measurement (in terms of composite reliability and average variance extracted) improves slightly.

4.2 Formative measurement models

It is not always meaningful to regard observed measures as reflective indicators of an underlying construct. Sometimes, observed measures are more properly thought of as antecedent influences on a construct of interest. One example often used to illustrate this idea is the concept

of socioeconomic status (SES) or social class. If one conceptualizes SES in terms of income, education, and occupational status, social class is probably not reflected in these variables, but more likely (at least partly) determined by them. As noted by MacKenzie *et al*. (2011) and Wilcox *et al*. (2008), constructs are not inherently reflective or formative; it depends on how a researcher attempts to measure them. For example, SES can also be measured as a reflective construct (e.g., as perceptions, either by the respondent or key informants, of a person's social class standing on such items as high-low social class or bottom-top of the social ladder).

We previously mentioned the criteria described by MacKenzie *et al*. (2005) to distinguish between reflective and formative measurement models. In this section, we first discuss important issues related to the specification of formative measurement models and then describe methods that can be used to assess the measurement quality of formative indicators.

4.2.1 Specification of formative measurement models

A formative measurement model can be specified as follows:

$$\eta = \sum_{i=1}^{I} \gamma_i \, y_i + \varsigma \tag{4.9}$$

where η is the formative construct, y_i is the i^{th} formative indicator ($i = 1, \ldots, I$), γ_i is the coefficient linking y_i to η, and ζ is an error term. Thus, η is a weighted linear combination of I observed variables, but the relationship is not exact because of the presence of the error term ζ. A specific formative measurement model with three formative indicators is depicted in Figure 4.5 (Panel B), which also shows a reflective measurement model with three reflective indicators for purposes of comparison (Panel A). The formative measurement model includes two reflective indicators; the reason for this will be explained below in the discussion of identification.

There are five important differences between the formative measurement model in Equation (4.9) and the reflective measurement model in Equation (4.1). First, as shown in Equation (4.9), in a formative measurement model constructs are a function of their indicators, whereas in a reflective measurement model each indicator is a function of the underlying construct(s). Second, in contrast to the conceptualization of constructs as latent common factors in reflective measurement models, constructs in formative measurement models are composite (or index) factors. The composite may be specified to contain no error, but this is usually not a reasonable assumption since it is unlikely that the formative indicators will completely capture the construct of interest. If the formative construct is assumed to contain error, as in Equation (4.9), it is called a latent composite.

Third, since reflective indicators have a common antecedent (i.e., the underlying factor), they are perforce positively correlated (assuming that the indicators are all scored in the same direction). In particular, since "good" reflective indicators are highly correlated with the underlying construct, the pairwise correlations between reflective indicators should be substantial. There is no such requirement for formative indicators. In fact, high correlations among formative indicators are undesirable because they may lead to multicollinearity problems.

Fourth, conditional on the common factors, the indicators in reflective measurement models are uncorrelated (assuming that the unique factors are uncorrelated), and if each indicator loads on a single factor, the measures of a given construct are unidimensional. In contrast, the indicators in formative measurement models are usually allowed to correlate freely, and in general the indicators are multi-dimensional. Some authors have argued that this makes it difficult to assign meaning to the resulting composite (Edwards, 2011). For example, in the context of social class, some researchers have suggested that it may be more meaningful to

investigate the separate relationships of income, education, and positional status with both antecedents and consequences of interest instead of creating an overall SES composite. An additional complication of allowing formative indicators to correlate freely is that formative measurement models may lack parsimony. For example, if there are three formative constructs with five indicators each, 105 pair-wise covariances between the indicators have to be estimated. Although it is possible to restrict some of these covariances to zero (e.g., only 30 covariances have to be estimated if the covariances between indicators of different constructs are restricted to zero), these restrictions are frequently not supported by the data and thus lead to poor model fit.

Fifth, in reflective measurement models, observed indicators are fallible manifestations of an underlying latent variable, which implies that measurement error resides in the observed variables. In contrast, formative indicators are often assumed to be error-free "measures" of the intended construct. Although this is an unrealistic assumption in most situations, it is possible to relax this assumption and specify multiple reflective indicators for each antecedent influence on the formative construct. For example, if satisfaction with one's salary, satisfaction with one's supervisor, and satisfaction with one's co-workers are used as formative indicators of the construct of job satisfaction, one could assess each of these facets of job satisfaction with multiple items, which makes it possible to take into account measurement error. One obvious disadvantage is that the resulting measurement model is rather complex and that it is no longer purely formative at the antecedent level. An alternative is to specify a single indicator for each antecedent influence on the formative construct, but instead of assuming the error variance associated with each formative indicator to be zero, the error variance is fixed to some other value that reflects the unreliability of measurement (e.g., based on reliabilities reported in previous studies or an estimate of reliability such as coefficient alpha when multiple items are

summed or averaged to create a composite indicator). While it is possible to model error in formative indicators, in most formative measurement models, error resides at the construct level, because the composite formed by the formative indicators has an error term. This error term does not capture measurement error directly (because, conventionally, measurement error afflicts observed measures), but to the extent that measurement error distorts the observed measures of the antecedent influences on the formative construct, the formative construct itself will be measured with error. Of course, there can also be other sources of error in the measurement of the formative construct, such as determinants of the construct that were not measured explicitly.

Although identification is important for any model, it is a particularly vexing problem for formative measurement models. In general, a model with a formatively measured construct is not identified. One way in which such a model can be identified is by assuming that at least two reflective indicators of an otherwise formatively measured construct are available (see Figure 4.5, Panel B). The resulting model is called a MIMIC (multiple-indicator multiple-cause) model. One problem with mixed (reflectively and formatively measured) constructs of this kind is that researchers usually use the minimum number of reflective measures necessary to identify the model (i.e., two) and that the two reflective measures are often not well-developed indicators of the assumed underlying construct. Furthermore, it is not clear whether the model should be interpreted as a measurement model containing a mix of formative and reflective indicators or as a reflective measurement model in which the construct is (possibly poorly) measured by two (or more) reflective indicators, and several antecedent variables (which are measured with single, supposedly error-free indicators) are hypothesized to explain the reflectively measured construct. Other methods to identify formative measurement models (besides using reflective indicators)

are available, but they are all problematic in various ways (see MacCallum and Brown, 1993, and Kline, 2013, for details).

As mentioned earlier, measurement models are sometimes misspecified (usually such that a reflective measurement model is assumed when a formative measurement model seems more appropriate). If the misspecification were innocuous, then researchers would not have to worry about using the correct measurement model. However, several papers indicate that measurement model misspecification does matter (see the review in Diamantopoulos *et al*., 2008). First, estimated structural paths between constructs may be biased. Second, if the item purification strategy used during measure development is inappropriate, the resulting scales might be poor. For example, if items are eliminated because of low internal consistency with other items even though the measures are formative (see below), the resulting scale may fail to capture essential facets of the formative construct.

4.2.2 Measurement analysis for formative indicator models

Assessing the quality of measurement for formative constructs is fundamentally different from the procedures discussed under reflective measurement models (see Diamantopoulos *et al*., 2008, and MacKenzie *et al*., 2011). Also, while the assessment of the reliability/convergent validity and discriminant validity of reflective measures is well-established, there is less agreement on how these aspects of measurement analysis apply to formative measures. Still, the following observations are probably reasonably non-controversial.

Since formative indicators need not be positively correlated, reliability based on internal consistency is not a meaningful concept. As an alternative, a researcher could assess reliability

based on the stability of construct scores over time (i.e., test-retest reliability). This assumes, of course, that the formative construct in question is stable over time.

The convergent validity of individual formative indicators is usually assessed based on the magnitude and significance of the γ_i parameters in Equation (4.9). A variation on this method is to compute the unique increment in the total explained construct variance contributed by a given formative indicator (i.e., the R^2 in the construct accounted for by all formative indicators minus the R^2 in the construct accounted for by the formative indicators excluding the one of interest). When the formative indicators are not too highly correlated, this is a useful diagnostic. However, even though there is no requirement that formative indicators should be highly correlated, in practice they often are, in which case shared variance among the indicators may make it difficult to discern the unique contribution of each individual formative indicator to the overall formative construct. Researchers are usually encouraged to use standard multicollinearity diagnostics to examine whether excessive shared variance is the culprit of lack of significance of individual γ_i parameters (e.g., variance inflation factors greater than 10 or tolerance less than .1), but these diagnostics are not always conclusive, and even if multicollinearity is found to be present, it is not clear whether the formative indicators are truly redundant (which implies that non-significant indicators can be dropped) or whether the redundancy is context-specific (which means that the indicator should not be dropped, because the meaning of a formative construct depends on its antecedent influences).

Some authors (e.g., MacKenzie *et al*., 2011) have suggested that when a formatively measured construct also has reflective indicators, then the correspondence between each formative indicator and each reflective indicator can be investigated. Essentially, this tests the strength of the indirect effect of a formative indicator on a reflective indicator (via the

formatively measured construct). However, since this method combines the convergent validity of formative and reflective indicators into one overall index, the diagnostic value of such a test is open to question.

It is also possible to assess the convergent validity of the entire set of formative indicators by investigating the proportion of the variance (R^2) in the formative construct accounted for by its indicators. If the R^2 is high, the formative indicators are able to capture a large portion of the variability in the latent composite. Of course, this is a meaningful index of convergent validity only in a pure measurement model, in which there are no other (substantive) determinants of the formative construct besides its formative indicators (otherwise the R^2 does not only assess the convergent validity of the set of formative indicators).

Discriminant validity at the item level can be assessed by specifying each formative indicator as an antecedent of both the target construct and the construct(s) from which the researcher tries to establish discriminant validity, or by looking at the modification indices for the paths from each formative indicator to non-target constructs. To assess discriminant validity at the construct level, a researcher can test whether the correlation between the target construct and related constructs is different from one, but as in the case of reflective measurement models, this may not be a strong test of discriminant validity. The so-called Fornell and Larcker criterion for assessing discriminant validity is not applicable (because the concept of average variance extracted is not meaningful), but it is possible to compare the squared correlation between constructs to the proportion of variance in the formative construct explained by its (formative) indicators.

There was only one paper among all the scale development efforts summarized in Table 2.1 in which observed measures were treated as formative indicators of the construct(s) of

interest. Reinartz *et al*. (2004) developed a scale for measuring the customer relationship management (CRM) process which consists of three higher-order dimensions (relationship initiation, maintenance, and termination), each of which comprises several sub-dimensions (9 in total), which are measured by multiple indicators. Both the relationship of the observed measures to the first-order sub-dimensions and the relationships between the sub-dimensions and higher-order dimensions were modeled formatively. Reinartz *et al*. conducted only a limited measurement analysis by reporting variance inflation factors for the indicators of each sub-dimension and used PLS to estimate a MIMIC model in which the implementation of CRM processes was also assessed with four reflective measures. They then formed composites for each of the three dimensions of the scale based on the standardized PLS weights and used the resulting indexes as independent variables in a regression analysis.

4.2.3 Additional issues related to formative measurement models

Although there has been substantial interest in formative measurement models in recent years, researchers' understanding of the complexities involved in these model is still evolving. Among the problematic features of formative measurement models are the following (e.g. Diamantopoulos, 2011; Diamantopoulos *et al*., 2008; Howell, Breivik, and Wilcox, 2007; Temme and Hildebrandt, 2006; Wilcox *et al*., 2008). First, the estimates based on formative indicator models may depend on the scaling of the formative construct (i.e., whether a formative or reflective indicator is used to set the scale of the formative construct or whether the variance of the latent composite is fixed to one, and which indicator is used as the reference indicator). Second, if a formative construct is specified as an outcome of other substantive constructs, in addition to being a consequence of its indicators, models in which these other constructs are only

related to the formative construct, not its indicators, may be misspecified (although this is not an issue in pure measurement models). Third, since a formative construct has to be related to at least two consequences in order for the model to be identified (either reflective indicators or other constructs), the meaning of the formative construct depends on which consequences are included in the model (called interpretational confounding). Because of all these difficulties, as well as others, some authors have argued that formative measurement models should be abandoned (Edwards, 2011; Howell, Breivik, and Wilcox, 2007).

Formative measurement models are closely associated with partial least squares (PLS) modeling in people's mind, even though formative measurement models can also be estimated using conventional structural equation modeling (SEM) techniques based on maximum likelihood and related estimation procedures and, conversely, PLS can also be applied to reflective measurement models. Although PLS has been popular in certain research domains (esp. in the marketing and information systems literatures), the technique has been strongly criticized in recent writings (see esp. Rönkkö *et al*. 2016). Among the major criticisms are that PLS is a methodologically deficient estimation algorithm (e.g., no sound justification is available for the way PLS weights are computed, which are needed to combine individual indicators into composites) and that the empirical support for some of its presumed advantages is weak (e.g., that PLS is preferred for small samples, non-normal data, and in the exploratory stages of research). Rönkkö *et al*. (2016, p. 24) conclude that "the only logical and reasonable action stemming from objective consideration of these issues is to discontinue the use of PLS and instead pursue superior alternatives, namely the ongoing stream of methodological innovations in latent variable-based SEM." Obviously, not everybody agrees with this view. For example, based on their simulation comparison of covariance-based SEM (CBSEM) and variance-based SEM

(PLS), Reinartz, Haenlein, and Henseler (2009, p. 332) conclude that "justifying the choice of PLS due to a lack of assumptions regarding indicator distribution and measurement scale is often inappropriate, as CBSEM proves extremely robust with respect to violations of its underlying distributional assumptions. Additionally, CBSEM clearly outperforms PLS in terms of parameter consistency and is preferable in terms of parameter accuracy as long as the sample size exceeds a certain threshold (250 observations). Nevertheless, PLS analysis should be preferred when the emphasis is on prediction and theory development, as the statistical power of PLS is always larger than or equal to that of CBSEM; already, 100 observations can be sufficient to achieve acceptable levels of statistical power given a certain quality of the measurement model."

One important issue that has not been discussed much in the context of formative measurement models is the assessment of measurement invariance (for exceptions see Diamantopoulos and Papadopoulos, 2010; Henseler, Ringle, and Sarstedt, 2016). The paper by Henseler *et al*. only deals with formative indicator models in which there is no error in the formative construct, and it is restricted to PLS estimation. Diamantopoulos and Papadopoulos (2010) propose three types of measurement invariance of formative measures in the context of international business research. Structure invariance means that the formative construct is determined by the same formative indicators in each country (i.e., that the same γ_i in equation (4.9) are non-zero in each country). Slope invariance means that the γ_i coefficients corresponding to the same formative indicators are the same in each country. Residual invariance means that the variance of the error associated with each formative construct is the same across countries. Diamantopoulos and Papadopoulos (2010) also suggest that partial measurement invariance may be sufficient (esp. for slope invariance), and they mention that since a "pure" formative measurement model is not identified, researchers usually use at least two reflective indicators as

additional measures of the formative construct, which should be "at least metrically invariant across countries" (p. 363). Finally, they suggest a three-step procedure according to which researchers should first test the metric invariance of the reflective indicators, then establish structure invariance using a baseline MIMIC model (in which the reflective indicators are specified to be metrically invariant), and finally test for slope and residual variance.

Diamantopoulos and Papadopoulos (2010) do not consider the means of the observed and latent variables, so their discussion is limited to the "loadings" of observed indicators on latent constructs. Since at least two reflective indicators must have invariant loadings, if researchers start out with only two indicators and then find that full metric invariance does not hold, further measurement invariance testing cannot proceed. Diamantopoulos and Papadopoulos start with a pure reflective measurement model consisting of three reflective indicators, establish metric invariance for that model, and then use only two of the three reflective indictors in the MIMIC model that also includes the formative indicators. It seems preferable to start with the full MIMIC model in which no constraints are imposed on any of the model parameters and use all available reflective indicators in the model. Researchers should then compare the following models to this baseline model: (a) a model in which the loadings of the reflective indicators are specified to be the same across groups (full metric invariance of the reflective indicators); (b) a model in which the γ_i coefficients of the formative indicators are specified to be the same across groups; and (c) a model in which both the loadings of the reflective indicators and γ_i coefficients of the formative indicators are specified to be invariant across groups. Chi-square difference tests can be used to compare these models, and if invariance of a given kind does not hold, modification indices can be used to free parameter constraints that are not supported by the data. At this point, little is known about how many γ_i coefficients should be invariant in order for

comparisons of structural relationships between constructs to be meaningful, but conceptually the notion of partial slope invariance is not very meaningful, because if important influences on the formative construct differ, the construct itself would seem to be noncomparable.

Even if full invariance of the loadings of the reflective indicators and the γ_i coefficients of the formative indicators is satisfied, there is a serious problem. The procedure assumes that the measurement of the formative indicators is invariant across groups. If only a single measure of each formative indicator is available, it is impossible to test whether this assumption is satisfied. If multiple reflective measures of each formative "indicator" are available, this assumption can be verified, but often this is not possible.

If a researcher wants to compare the means of formative constructs across groups, it is necessary to specify a model for the mean structure, which includes the means of the observed variables (both the reflective and formative indicators), intercepts for the reflective indicators, and intercepts for the formative constructs. The means of the formative constructs are functions of the γ_i coefficients, including an intercept term that has to be added to equation (4.9), and the means of the observed formative indicators. It is straightforward to test for the metric and scalar invariance of the reflective indicators, but even if the invariance of the γ_i coefficients (including the intercept) can be established, the invariance of the measurement of the formative indicators (which is necessary for the means of the formative indicators to be comparable across groups) remains an unverified assumption.

In summary, although we agree that a reflective measurement model is not appropriate for some indicators, there are so many problems with formative measurement models that it is difficult to recommend their use except in special circumstances.

5

Conclusion

Measurement is a multi-faceted and intricate activity involving both research design issues (described in sections 2 and 3) and data analysis issues (described in section 4). The former requires skills in conceptualizing constructs and developing observed indicators of these constructs, as well as an intricate understanding of how respondents react to the questions they are being asked and how they generate an observed response. The latter requires an ability to specify and test possibly complex measurement models in an effort to ascertain whether the questions designed to capture the researcher's constructs were successful in getting respondents to provide valid and reliable responses.

In this monograph we described what we consider to be the most important issues facing a researcher who wants to measure constructs of interest. Researchers can use the discussion in sections 2 to 4 as a checklist to remind them of the things they should consider when designing and analyzing the measurement aspects of their research. Although measurement is usually not the primary concern of empirical researchers who are interested in substantive topics, unless measurement is done well, the substantive findings are of questionable value. It has become very easy (at least in some domains of research) to collect data quickly and inexpensively, but unfortunately the measures that are used to capture constructs of interest are often poorly designed, respondents are frequently not sufficiently motivated and possibly not able to provide the desired responses, and the data collected from respondents are analyzed after performing only the most rudimentary measurement analysis. It is our hope that this monograph will alert

researchers to the importance of measurement in the research process and provide them with tangible advice on what they can do to improve their measurement practices.

References

Arce-Ferrer, A. J. (2006). "An investigation into the factors influencing extreme-response style". *Educational and Psychological Measurement, 66* (3): 374-392.

Bagozzi, R. P. (1980). *Causal models in marketing*. New York: Wiley.

Bagozzi, R. P. (1984). "A prospectus for theory construction in marketing". *Journal of Marketing*, 48 (1): 11-29.

Bagozzi, R. P., and L. W. Phillips (1991). "Assessing constuct validity in organizational research". *Administrative Science Quarterly*, 36 (September): 421-458.

Baka, A., L. Figgou, and V. Triga (2012). "'Neither agree, nor disagree': A critical analysis of the middle answer category in voting advice applications". *International Journal of Electronic Governance*, 5 (3): 244-63.

Bandalos, D. L. (2020). "The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling". *Structural Equation Modeling*, 9(1): 78-102.

Batra, R., A. Ahuvia, and R. P. Bagozzi (2012). "Brand love". *Journal of Marketing*, 76 (March): 1-16.

Baumgartner, H. and J.-B. E.M. Steenkamp (2001). "Response styles in marketing research: A cross-national investigation". *Journal of Marketing Research*, 38 (May): 143-156.

Baumgartner, H. and B. Weijters (2012). "Commentary on "Common method bias in marketing: Causes, mechanisms, and procedural remedies". *Journal of Retailing*, 88 (4): 563-566.

Baumgartner, H., and B. Weijters (2015). "Response biases in cross-cultural measurement". In: *Handbook of Culture and Consumer Behavior.* Ed. by S. Ng and A. Y. Lee, Oxford, UK: Oxford University Press. 150-180.

Baumgartner, H. and B. Weijters (2017). "Measurement models for marketing constructs". In: *Handbook of Marketing Decision Models*, 2nd ed., Ed. by B. Wierenga and R. van der Lans. Cham, Switzerland: Springer International Publishing AG. 259-295.

Baumgartner, H., B. Weijters, and R. Pieters (2018). "Misresponse to survey questions: A conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts". *Journal of Marketing Research*, 55 (6): 869-883.

Bearden, W. O., D. M. Hardesty, and R. L. Rose (2001). "Consumer self-confidence: Refinements in conceptualization and measurement". *Journal of Consumer Research*, 28 (June): 121-134.

Bearden, W. O., R. G. Netemeyer, and K. Haws (2011). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research*, 3rd ed., Palo Alto, CA: Sage Publications.

Bergkvist, L., and J. R. Rossiter (2007). "The predictive validity of multi-item versus single-item measures of the same constructs". *Journal of Marketing Research*, 44 (May): 175-184.

Bettman, J. R. (1979), *An information processing theory of consumer choice*, Reading, MA: Addison-Wesley.

Bettman, J. R., M. F. Luce, and J. W. Payne (1998). "Constructive consumer choice processes". *Journal of Consumer Research*, 25 (December): 187-217.

Bloch, P. H., F. F. Brunel, and T. J. Arnold (2003). "Individual differences in the centrality of visual product aesthetics: Concept and measurement". *Journal of Consumer Research*, 29 (March): 551-565.

Böckenholt, U. (2017). "Measuring response styles in Likert items". *Psychological Methods*, 22 (March), 69-83.

Bollen, K. A. and K. H. Barb (1981). "Pearson's R and coarsely categorized measures". *American Sociological Review*, 46 (April): 232-39.

Bolt, D. M, Y. Lu, and J.-S. Kim (2014). "Measurement and control of response styles using anchoring vignettes: A model-based approach". *Psychological Methods*, 19 (4): 528-541.

Böttger, T., T. Rudolph, H. Evanschitzky, and T. Pfrang (2017). "Customer inspiration: Conceptualization, scale development, and validation." Journal of Marketing, 81 (November): 116-131.

Brady, M. K., and J. Cronin Jr. (2001). "Some new thoughts on conceptualizing perceived service quality: A hierarchical approach". *Journal of Marketing*, 65 (July): 34-49.

Bruner II, G. C. (2015). *Marketing scales handbook*. GCBII Productions, LLC.

Cabooter, E., B. Weijters, M. Geuens, and I. Vermeir (2016). "Scale format effects on response option interpretation and use". Journal of Business Research, 69 (7), 2574-2584.

Cabooter, E., K. Millet, and B. Weijters (2016). "The 'I' in extreme responding". *Journal of Consumer Psychology,* 26 (4): 510-523.

Campbell, D. T., and D. W. Fiske (1959). "Convergent and discriminant validity by the multitrait-multimethod matrix". *Psychological Bulletin*, 56: 81-105.

Churchill, G. A., Jr. (1979). "A paradigm for developing better measures of marketing constructs". *Journal of Marketing Research*, 16 (February): 64-73.

Cole, D. A, C. E Perkins, and R. L Zelkowitz (2016). "Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs". *Psychological Methods*, 21 (2): 164-174.

Conijn, J. M., W. H. M. Emons, and M. A. L. M. van Assen (2013). "Explanatory, multilevel person-fit analysis of response consistency on the Spielberger Statet-Trait Anxiety Inventory". *Multivariate Behavioral Research, 48* (5): 692-718.

Converse, J. M. and S. Presser (1986), *Survey questions: Handcrafting the standardized questionnaire*, Beverly Hills: Sage Publications.

Cooper, W. H. (1981). "Ubiquitous halo". *Psychological Bulletin*, 90 (2): 218-244.

Cox III, E. P (1980). "The optimal number of response alternatives for a scale: A review". *Journal of Marketing Research*, 19 (November): 407-22.

Crowne, D. P. and D. Marlowe (1964). *The approval motive*. New York: Wiley.

Davies, M. F. (2003). "Confirmatory bias in the evaluation of personality descriptions: Positive test strategies and output interference." *Journal of Personality and Social Psychology*, 85 (4): 736–744.

De Jong, M. G., J.-B. E. M. Steenkamp, and J.-P. Fox (2007). "Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model". *Journal of Consumer Research*, 34 (22): 260-78.

DeCastellarnau, A. (2018). "A classification of response scale characteristics that affect data quality: A literature review". *Quality and Quantity*, 52: 123-1559.

DeVellis, R. F. (2003), *Scale development: Theory and application*, 2nd ed., Thousand Oaks, CA: Sage.

Diamantopoulos, A. (2011). "Incorporating formative measures into covariance-based structural equation models". *MIS Quarterly*, 35 (June): 335-358.

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

Diamantopoulos, A., and N. Papadopoulos (2010). "Assessing the cross-national invariance of formative measures: Guidelines for international business researchers". *Journal of International Business Studies*, 41 (February-March): 360-370.

Dixon, A. L., R. L. Spiro, and M. Jamil (2001). "Successful and unsuccessful sales calls: Measuring salesperson attributions and behavioral intentions". *Journal of Marketing*, 65 (July): 64-78.

Edwards, J. R. (2011). "The fallacy of formative measurement". *Organizational Research Methods,* 14 (2): 370-388.

Eid, M. (2000). "A multitrait-multimethod model with minimal assumptions". Psychometrika, 65 (2): 241-261.

Emons, W. H. M., K. Sijtsma, and R. R. Meijer (2005). "Global, local, and graphical person-fit analysis using person-response functions". *Psychological Methods, 10* (1): 101-119.

Ferrando, P. J., and U. Lorenzo-Seva (2007). "A measurement model for Likert responses that incorporates response time". *Multivariate Behavioral Research, 42* (4): 675-706.

Fiske, S. T. and S. E. Taylor (1991), *Social cognition,* 2nd ed. Reading, MA: Addison-Wesley Publishing Company.

Fornell, C. and D. F. Larcker (1981). "Evaluating structural equation models with unobservable variables and measurement error". *Journal of Marketing Research, 18* (1): 39-50.

Garner, W. R. (1960). "Rating scales, discriminability, and information transmission". *Psychological Review*, 67 (6): 343-352.

Gehlbach, H. amd S. Barge (2012). "Anchoring and adjusting in questionnaire responses". *Basic and Applied Social Psychology, 34:* 417-433.

Graesser, A. C., Z. Cai, M. M. Louwerse, and F. Daniel (2006). "Question understanding aid (QUAID): A web facility that tests question comprehensibility". *Public Opinion Quarterly, 70* (1): 3-22.

Graesser, A. C., K. Wiemer-Hastings, R. Kreuz, P. Wiemer-Hastings, and K. Marquis (2000). "QUAID: A questionnaire evaluation aid for survey methodologists". *Behavior Research Methods, Instruments, & Computers, 32* (2): 254-262.

Green, P. E and V. R Rao (1970). "Rating scales and information recovery—How many scales and response categories to use?" *Journal of Marketing*, 34 (July): 33-39.

Greenleaf, E. A. (1992a). "Improving rating scale measures by detecting and correcting bias components in some response styles". *Journal of Marketing Research*, 29 (2): 176-188.

Greenleaf, E. A. (1992b). "Measuring extreme response style". *Public Opinion Quarterly*, 56 (3): 328-350.

Hauser, D. J. and N. Schwarz (2016). "Attentive Turkers: Mturk participants perform better on online attention checks than do subject pool participants". *Behavior Research Methods*, 48 (1): 400-07.

Grohmann, B. (2009). "Gender dimensions of brand personality". *Journal of Marketing Research*, 46 (February): 105-119.

Groves, R. M., F. J. J. Fowler, M. P. Cooper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Hardy, B. and L. R. Ford (2014). "It's not me, it's you: Miscomprehension in surveys". *Organizational Research Methods*, 17 (2): 138-162.

Harrison, D. A. and M. E. McLaughlin (1993). "Cognitive processes in self-report responses: Tests of item context effects in work attitude measures". *Journal of Applied Psychology, 78* (1): 129-140.

Harrison, D. A., and M. E. McLaughlin (1996). "Structural properties and psychometric qualities of organizational self-reports: Field tests of connections predicted by cognitive theory". *Journal of Management, 22* (2): 313-338.

Hastie, R. and B. Park (1986). "The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line". *Psychological Review*, 93 (3): 258-268.

Hauser, D. J., and N. Schwarz (2015). "It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks". SAGE Open, 5 (April-June): 1-6.

Hauser, D. J., A. Sunderrajan, M. Natarajan, and N. Schwarz (2016). "Prior exposure to instructional manipulation checks does not attenuate survey context effects driven by satisficing or Gricean norms". *Methods, data, analyses: A journal for quantitative methods and survey methodology*, 10 (2): 195-220.

Henseler, J., C. M. Ringle, and M. Sarstedt (2016). "Testing measurement invariance of composites using partial least squares". *International Marketing Review*, 33 (3): 405-431.

Holt, J. K. (2004). "Item parceling in structural equation models for optimum solutions. Paper presented at the 2004 Annual Meeting of the Mid-Western Educational Research Association, October 13–16, 2004, Columbus, OH.

Homburg, C. and C. Pflesser (2000). "A multiple-layer model of market-oriented organizational culture: Measurement issues and performance outcomes". *Journal of Marketing Research*, 37 (November): 449-462.

Homburg, C., M. Schwemmle, and C. Kuehnl (2015). "New product design: Concept, measurement, and consequences". *Journal of Marketing*, 79 (May): 41-56.

Hovland, C. I., I. K. Janis, and H. H. Kelley (1953). *Communication and persuasion*. New Haven, CT: Yale University Press.

Howell, R. D., E. Breivik, and J. B. Wilcox (2007). "Reconsidering formative measurement". *Pscyhological Methods*, 12 (2): 205-218.

Hsee, C. K., Y. Zang, X. Zheng, and H. Wang (2015). "Lay rationalism: Individual differences in using reason versus feelings to guide decisions". *Journal of Marketing Research*, 52 (February): 134-146.

Huang, J. L., P. G. Curran, J. Keeney, E. M. Poposki, and R. P. DeShon (2012). "Detecting and deterring insufficient effort responding to surveys". *Journal of Business and Psychology, 27:* 99-114.

Hui, C. H. and H. C. Triandis (1985). "The instability of response sets". *Public Opinion Quarterly*, 49: 253-260.

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

Johnson, J. A. (2005). "Ascertaining the validity of individual protocols from Web-based personality inventories". *Journal of Research in Personality, 39:* 103-129.

Judd, C. M., and G. H. McClelland (1998). "Measurement". In: *The handbook of social psychology,* 4th ed. Ed. by D. Gilbert, S. T. Fiske, and G. Lindzey. New York, NY: McGraw-Hill. 180-232.

Kam, C. C. S. and G. H. Chan (2018). "Examination of the validity of instructed response items in identifying careless respondents". *Personality and Individual Differences*, 129, 83-87.

Kamakura, W. A. (2015). "Measure twice and cut once: The carpenter's rule still applies". *Marketing Letters*, 26: 237-243.

Kamata, A. and D. .J. Bauer (2008). "A note on the relation between factor analytic and item response theory models". *Structural Equation Modeling, 15* (1): 136-153.

Kidwell, B., D. M. Hardesty, and T. L. Childers (2008). "Consumer emotional intelligence: Conceptualization, measurement, and the prediction of consumer decision making". *Journal of Consumer Research*, 35 (June): 154-166.

King, G. and J. Wand (2007). "Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes". *Political Analysis*, 15 (1): 46-66.

Kline, R. B. (2013). “Reverse arrow dynamics: Feedback loops and formative measurement”. In: *Structural Equation Modeling: A Second Course*, 2nd ed. Ed. by G. R. Hancock and R. O. Mueller, Greenwich, CT: Information Age Publishing. 39-76.

Krosnick, J. A. (1991). “Response strategies for coping with the cognitive demands of attitude measures in surveys”. *Applied Cognitive Psychology, 5:* 213-236.

Krosnick, J. A. and L. R. Fabrigar (1997). “Designing rating scales for effective measurement in surveys”. In: Survey measurement and process quality. Ed. by L. Lyberg, P. Biemer, M.Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York, NY: John Wiley & Sons. 141-164.

Kunda, Z., G. T. Fong, R. Santioso, and E. Reber (1993). “Directional questions direct self-conceptions”. *Journal of Experimental Social Psychology*, 29: 63-86.

Lance, C. E., J. A. LaPointe, and S. A. Fisicaro (1994). “Test of three causal models of halo rater error”. *Organizational Behavior and Human Decision Processes*, 57: 83-96.

Lance, C. E., C. L. Noble, and S. E. Scullen (2002). “A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data”. *Psychological Methods, 7* (2): 228-244.

Lastovicka, J. and N. J. Sirianni (2011). “Truly, madly, deeply: Consumers in the throes of material possession love”. *Journal of Consumer Research*, 38 (August): 323-339.

Lenzner, T. (2012). “Effects of survey question comprehensibility on response quality”. *Field Methods, 24* (4): 409-428.

Lenzner, T. (2014). “Are readability formulas valid tools for assessing survey question difficulty?” *Sociological Methods & Research*, 43 (4): 677-698.

Lenzner, T., L. Kaczmirek, and M. Galesic (2011). “Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension”. *International Journal of Public Opinion Research*, 23 (3): 361-373.

Lenzner, T., L. Kaczmirek, and A. Lenzner (2010). “Cognitive burden of survey questions and response times: A psycholinguistic experiment”. *Applied Cognitive Psychology, 24* (7): 1003-1020.

Lynch, J. G., Jr., R. G. Netemeyer, S. A. Spiller, and A. Zammit (2009). “A generalized scale of propensity to plan: The long and the short of planning for time and for money”. *Journal of Consumer Research*, 37 (June): 108-128.

MacCallum, R. C. (1986). “Specification searches in covariance structure modeling”. *Psychological Bulletin*, 100 (1): 107-120.

MacCallum, R. C. and M. W. Browne (1993). “The use of causal indicators in covariance structure models: Some practical issues”. *Psychological Bulletin*, 114: 533-541.

MacCallum, R. C., M. Roznowski, and L. B. Necowitz (1992). “Model modification in covariance structure analysis: The problem of capitalization on chance”. *Psychological Bulletin*, 111 (3): 490-504.

MacKenzie, S. B. and P. M. Podsakoff (2012). "Common method bias in marketing: Causes, mechanisms, and procedural remedies". *Journal of Retailing, 88* (4): 542-555.

MacKenzie, S. B., P. M. Podsakoff, and C. B. Jarvis (2005). "The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions". *Journal of Applied Psychology, 90* (4): 710-730.

MacKenzie, S. B., P. M. Podsakoff, and N. P. Podsakoff (2011). "Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques". *MIS Quarterly*, 35 (June): 293-334.

Marsh, H. W. (1989). "Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions". *Applied Psychological Measurement*, 14 (4): 335-361.

Marsh, H. W., A. J. Morin, P. D. Parker, and G. Kaur (2014). "Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis". *Annual Review of Clinical Psychology, 10:* 85-110.

Maydeu-Olivares, A. and D. L. Coffman (2006). "Random intercept item factor analysis". *Psychological Methods*, 11 (4): 344–362.

Meade, A. W. and S. B. Craig (2012). "Identifying careless responses in survey data. *Psychological Methods*, *17* (3): 437-455.

Meijer, R. R. (2003). "Diagnosing item score patterns on a test using item response theory-based person-fit statistics". *Psychological Methods*, 8 (1): 72-87.

Millsap, R. E. and J. Yun-Tein (2004). "Assessing factorial invariance in ordered-categorical measures". *Multivariate Behavioral Research*, 39 (3): 479-515.

Moors, G., N. D Kieruj, and J. K Vermunt (2014). "The effect of labeling and numbering of response scales on the likelihood of response bias". *Sociological Methodology*, 44 (1): 369-399.

Morren, M., J. P. T. M. Gelissen, and J. K. Vermunt (2011). "Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach". *Sociological Methodology*, 41: 13-47.

Muraki, E. (1990). "Fitting a polytomous item response model to Likert-type data". *Applied Psychological Measurement, 14* (1): 59-71.

Muthén, B. and T. Asparouhov (2012). "Bayesian structural equation modeling: A more flexible representation of substantive theory". *Psychological Methods, 17* (3): 313-335.

Muthén, B. and T. Asparouhov (2018). "Recent methods for the study of measurement invariance with many groups: Alignmet and random effects." *Sociological Methods & Research, 47* (4): 737-664.

Nadler, J. T., R. Weston, and E. C. Voyles (2015). "Stuck in the middle: The use and interpretation of mid-points in items on questionnaires". *The Journal of General Psychology*, 142 (2): 71-89.

Nowlis, S. M., B. E. Kahn, and R. Dhar (2002). "Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments". *Journal of Consumer Research*, 29: 319-34.

Nenkov, G. Y., J. J. Inman, and J. Hulland (2008). "Considering the future: The conceptualization and measurement of elaboration on potential outcomes". *Journal of Consumer Research*, 35 (June): 126-141.

Netemeyer, R. G., W. O. Bearden, and S. Sharma (2003). *Scaling procedures: Issues and applications*, Thousand Oaks, CA: Sage.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Oppenheimer, D. M., T. Meyvis, and N. Davidenko (2009). "Instructional manipulation checks: Detecting satisficing to increase statistical power". *Journal of Experimental Social Psychology, 45:* 867-872.

Parasuraman, A., V. A. Zeithaml, and L. L. Berry (1988). "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality". *Journal of Retailing,* 64 (Spring): 12–40.

Paulhus, D. L. (1991). "Measurement and control of response bias". In: *Measures of personality and social psychological attitudes.* Ed. by J. P. Robinson, P. R. Shaver, and L. S. Wright. San Diego, CA: Academic Press. 17-59.

Paulhus, D. L. (2002). "Socially desirable responding: The evolution of a construct". In: *The role of constructs in psychological and educational measurement.* Ed. by H. I. Braun, D.N. Jackson, and D.E. Wiley. Mahwah, NJ: Erlbaum. 49-69.

Peck, J. and T. L. Childers (2003). "Individual differences in haptic information processing: The "Need for Touch" scale". *Journal of Consumer Research*, 30 (December): 430-442.

Peter, J. P. (1981). "Construct validity: A review of basic issues and marketing practices". *Journal of Marketing Research*, 18 (May): 133-145.

Petty, R. E., J. T. Cacioppo, A. J. Strathman, and J. Priester (2005). "To think or not to think? Exploring two routes to persuasion," In: *Persuasion: Psychological insights and perspectives*, 2nd ed. Ed. by T. C. Brock and M. C. Green. Thousand Oaks: CA: Sage Publications. 81-116.

Podsakoff, P. M., S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff (2003). "Common method biases in behavioral research: A critical review of the literature and recommended remedies". *Journal of Applied Psychology*, 88 (5): 879-903.

Pornpitakpan, C. (2004). "The persuasiveness of source credibility: A critical review of five decades' evidence". *Journal of Applied Social Psychology*, 34 (2): 243-281.

Preston, C. C and A. M. Colman (2000). "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences". *Acta Psychologica*, 104 (1): 1-15.

Puligadda, S., W. T. Ross Jr., and R. Grewal (2012). "Individual differences in brand schematicity". *Journal of Marketing Research*, 49 (February): 115-130.

Reich, B. J., J. T. Beck, J. Price, & C. Lamberton (2018). Food as Ideology: Measurement and Validation of Locavorism. Journal of Consumer Research.

Reinartz, W., M. Haenlein, and J. Henseler (2009). "An empirical comparison of the efficacy of covariance-based and variance-based SEM". *International Journal of Research in Marketing*, 26 (4): 332-344.

Reinartz, W., M. Krafft, and W. D. Hoyer (2004). "The customer relationship management process: Its measurement and impact on performance". *Journal of Marketing Research*, 41 (August): 293-305.

Reise, S. P. and K. F. Widaman (1999). "Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches". *Psychological Methods, 4* (1): 3-21.

Rhemtulla, M., A. E. Brosseau-Liard, and V. Savalei (2012). "When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions". *Psychological Methods*, 17 (3): 354-373.

Richins, M. L. (2004). "The material values scale: Measurement properties and development of a short form". *Journal of Consumer Research*, 31 (June): 209-219.

Richins, M. L. and S. Dawson (1992). "A consumer values orientation for materialism and its measurement: Scale development and validation". *Journal of Consumer Research*, 19 (December): 303-316.

Rönkkö, M., C. N McIntosh, J. Antonakis, and J. R. Edwards (2016). "Partial least squares path modeling: Time for some serious second thoughts". *Journal of Operations Management*, 47-48 (November), 9-27.

Rick, S., I., C. E. Cryder, and G. Loewenstein (2008). "Tightwads and spendthrifts". *Journal of Consumer Research*, 34 (April): 767-782.

Ridgway, N. M., M. Kukar-Kinney, and K. B. Monroe (2008). "An expanded conceptualization and a new measure of compulsive buying". *Journal of Consumer Research*, 35 (December): 622-639.

Rossiter, J. R. (2002). "The C-OAR-SE procedure for scale development in marketing". *International Journal of Research in Marketing*, 19 (December): 305-335.

Russell, C. A., A. T. Norman, and S. E. Heckler (2004). "The consumption of television programming: Development and validation of the connectedness scale". *Journal of Consumer Research*, 31 (June): 150-161.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph No. 17,* Richmond, VA: Psychometric Society.

Saris, W., M. Revilla, J. A. Krosnick, and E. M. Shaeffer (2010). "Comparing questions with agree/disagree response options to questions with item-specific response options," *Survey Research Methods*, 4 (May): 61-79.

Schacter, D. L. (1999). "The seven sins of memory". *American Psychologist*, 54 (March): 182-203.

Schwarz, N. (1999). "Self-reports: How the questions shape the answers". *American Psychologist, 54:* 93-105.

Schwarz, N., B. Knäuper, H.-J. Hippler, E. Noelle-Neumann, and L. Clark (1991). "Rating scales: Numeric values may change the meaning of scale labels". *Public Opinion Quarterly*, 55: 570-582.

Schwarz, N., F. Strack, G. Mueller, and B. Chassein (1988). "The range of response alternatives may determine the meaning of a question: Further evidence on informative functions of response alternatives". *Social Cognition*, 6 (2): 107-117.

Sprott, D., S. Czellar, and E. Spangenberg (2009). "The importance of a general measure of brand engagement on market behavior: Development and validation of a scale". *Journal of Marketing Research*, 46 (February): 92-104.

Srinivasan, V. and A. K. Basu (1989). "The metric quality of ordered categorical data". *Marketing Science*, 8 (3): 205-30.

Steenkamp, J.-B. E.M. and H. Baumgartner (1998). "Assessing measurement invariance in cross-national consumer research". *Journal of Consumer Research*, 25 (June): 78-90.

Steenkamp, J.-B. E.M., M. G. De Jong, and H. Baumgartner (2010). "Socially desirable response tendencies in survey research". *Journal of Marketing Research*, 47 (April): 199-214.

Steenkamp, J.-B. E.M. and H. C. M. van Trijp (1991). "The use of LISREL in validating marketing constructs". *International Journal of Research in Marketing*, 8: 283-299.

Sterba, S. K. (2011). "Implications of parcel-allocation variability for comparing fit of item-solutions and parcel-solutions". *Structural Equation Modeling*, 18 (4): 554-577.

Sterba, S. K. and J. Pek (2012). "Individual influence on model selection". *Psychological Methods*, 17 (4): 582-599.

Sterba, S. K., and J. D. Rights (2017). "Effects of parceling on model selection: Parcel-allocation variability in model ranking". *Psychological Methods*: 22 (1), 47-68.

Stevens, S. S. (1946). "On the theory of scales of measurement". *Science*, 103 (June): 677-680.

Strack, F., N. Schwarz, and M. Wänke (1991). "Semantic and pragmatic aspects of context effects in social and psychological research". *Social Cognition*, 9 (1): 111-125.

Temme, D. and L. Hildebrandt (2006). "Formative measurement models in covariance structure analysis: Specification and identification". SFB 649 Discussion Paper 2006-083, Humboldt University Berlin.

Tian, K. T., W. O. Bearden, and G. L. Hunter (2001). "Consumers' need for uniqueness: Scale development and validation". *Journal of Consumer Research*, 28 (June): 50-66.

Tourangeau, R., L. J. Rips, and K. A. Rasinski (2000). *The psychology of survey response.* Cambridge University Press.

Tourangeau, R. and T. Yan (2007). "Sensitive questions in surveys". *Psychological Bulletin*, 133 (5): 859-883.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage.

Vandenberg, R. J. and C. E. Lance (2000). "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research". *Organizational Research Methods*, 3 (January): 4-69.

Viswanathan, M., X. Li, G. John, and O. Narasimhan (2018). "Is cash king for sales compensation plans? Evidence from a large-scale field intervention". *Journal of Marketing Research*, 55 (3): 368-381.

Voss, K. E., E. R. Spangenberg, and B. Grohmann (2003). "Measuring the hedonic and utilitarian dimensions of consumer attitude". *Journal of Marketing Research*, 40 (August): 310-320.

Weijters, B. and H. Baumgartner (2012). "Misresponse to reversed and negated items in surveys: A Review". *Journal of Marketing Research, 49* (5): 737-747.

Weijters, B., H. Baumgartner, and M. Geuens (2016). "The calibrated sigma method: An efficient remedy for between-group differences in response category use on Likert scales". *International Journal of Research in Marketing*, 33 (4): 944-60.

Weijters, B., H. Baumgartner, and N. Schillewaert (2013). "Reversed item bias: An integrative model". *Psychological Methods*, 18 (September): 320-334.

Weijters, B., Cabooter, E., and N. Schillewaert (2010). "The effect of rating scale format on response styles: The number of response categories and response category labels". *International Journal of Research in Marketing*, 27 (September), 236-247.

Weijters, B., A. De Beuckelaer, and H. Baumgartner (2014). "Discriminant validity where there should be none: Positioning same-scale items in separated blocks of a questionnaire". *Applied Psychological Measurement*, 38 (6): 450-463.

Weijters, B., M. Geuens, and N. Schillewaert (2009). "The proximity effect: The role of inter-item distance on reverse-item bias". *International Journal of Research in Marketing*, 26: 2-12.

Weijters, B., M. Geuens, and N. Schillewaert (2010a). "The individual consistency of acquiescence and extreme response style in self-report questionnaires". *Applied Psychological Measurement, 34* (2): 105-121.

Weijters, B., M. Geuens, and N. Schillewaert (2010b). "The stability of individual response styles". *Psychological Methods, 15* (1): 96-110.

Weijters, B., S. Puntoni, and H. Baumgartner (2017). "Methodological issues in cross-linguistic and multilingual advertising research". *Journal of Advertising*, 46 (1): 115-128.

Weijters, B., N. Schillewaert, and M. Geuens (2008). "Assessing response styles across modes of data collection". *Journal of the Academy of Marketing Science*, 36 (3): 409–22.

Weng, L.-J. (2004). "Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability". *Educational and Psychological Measurement*, 64 (6): 956-972.

Widaman, K. F. (1985). "Hierarchically nested covariance structure models for multitrait-multimethod data". *Applied Psychological Measurement*, 9 (1): 1-26.

Wilcox, J. B., R. D. Howell, and E. Breivik (2008). “Questions about formative measurement”. *Journal of Business Research*, 61: 1219-1228.

Wildt, A. R. and M. B. Mazis (1978). “Determinants of scale response: Label versus position”. *Journal of Marketing Research*, 15 (May): 261-67.

Wilkie, W. L. and E. A. Pessemier (1973). “Issues in marketing’s use of multi-attribute attitude models”. *Journal of Marketing Research*, 10 (November): 428-441.

Wise, S. L. And X. Kong (2005). “Response time effort: A new measure of examinee motivation in computer-based tests”. *Applied Measurement in Education, 18* (2): 163-183.

Wu, H. and R. Estabrook (2016). “Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes”. *Psychometrika*, 81 (4): 1014-1045.

Zettler, I., J. W. B. Lang, U. R. Hülsheger, and B. E. Hilbig (2015). “Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self-and observer reports”. *Journal of Personality*, 84 (4): 461-472.

Zou, S. and S. T. Cavusgil (2002). “The GMS: A broad conceptualization of global marketing strategy and its effect on firm performance”. *Journal of Marketing*, 66 (October): 40-56.

Table 2.1: Illustrative scale development papers published in the *Journal of Consumer Research*, *Journal of Marketing*, and *Journal of Marketing Research* since 2000

Construct	Definition	Dimensions of the construct	Measures	Measurement model	Object and attribute of quantification as well as rater
Market-oriented organization culture (MOOC) (Homburg and Pflesser, 2000)	No explicit definition provided.	Four dimensions: organization-wide shared basic values supporting market orientation, organization-wide norms for market orientation, perceptible artifacts of market orientation, and market-oriented behaviors. Values are the most basic layer, norms are the next layer, and artifacts and behaviors form the third layer.	Eight dimensions of values and norms each (e.g., success, speed, and innovativeness, measured by a total of 22 and 25 items, respectively), 6 dimensions of artifacts (e.g., stories, rituals, language, 19 items), and one dimension of behaviors (12 items), rated on 7-point strongly agree-disagree or frequency scales. Two artifact dimensions indicate lack of market orientation, but no real reversed items.	Initially, a reflective correlated 23-factor model was estimated, but eventually a 5-factor model was used (values, norms, artifacts indicating the presence or absence of market orientation, and behaviors). The final model specifies causal paths between the different layers of market orientation.	Quantification of an SBU (object) on MOOC (attribute) through key-informant reports (rater).
Consumer need for uniqueness (CNFU) (Tian, Bearden, and Hunter, 2001)	CNFU as "the trait of pursuing differentness relative to others through the acquisition, utilization, and disposition of consumer goods for the purpose of developing and enhancing one's self-image and social image" (p. 52).	Three dimensions: creative choice counterconformity (CCC), unpopular choice counterconformity (UCC), and avoidance of similarity (AS)	I actively seek to develop my personal uniqueness by buying special products or brands (CCC, 11 items), I have often gone against the understood rules of my social group regarding when and how certain products are properly used (UCC, 11 items), I often try to avoid products or brands that I know are bought by the general population (AS, 9	Reflective correlated three-factor model.	Quantification of a consumer (object) on CNFU (attribute) through self-report (rater).

			items), rated on 5-point strongly agree/disagree scale. No reversed items.		
Consumer self-confidence (CSC) (Bearden, Hardesty, and Rose, 2001)	CSC as "the extent to which an individual feels capable and assured with respect to his or her marketplace decisions and behaviors" (p. 122).	Two higher-order dimensions (decision-making self-confidence or DM, protection or PROT), each made up of multiple subdimensions (information acquisition or IA, consideration set formation or CSF, personal outcomes or PO, and social outcomes or SO for DM; persuasion knowledge or PK and marketplace interfaces or MI for PROT).	I know where to look to find the product information I need (IA, 5 items), I am confident in my ability to recognize a brand worth considering (CSF, 5 items), I often have doubts about the purchase decisions I make (PO, 5 items), I impress people with the purchases I make (SO, 5 items), I can see through sales gimmicks used to get consumers to buy (PK, 6 items), I am afraid to "ask to speak to the manager" (MI, 5 items), rated on 5-point characteristic scales. All items for two of the factors (PO, MI) are reversed.	A reflective correlated six-factor model yielded the best fit, although a reflective second-order factor model with two factors was also investigated.	Quantification of a consumer (object) on CSC (attribute) through self-report (rater).
Perceived service quality (PSQ) (Brady and Cronin, 2001)	No explicit definition provided, but according to the authors, "customers form their service quality perceptions on the basis of an evaluation of performance at multiple levels and ultimately combine these evaluations to arrive at an overall service quality perception" (p. 37).	Three primary dimensions: interaction quality (IQ), physical environmental quality (PEQ), and outcome quality (OQ), each of which consists of three subdimensions: attitude, behavior, and expertise for IQ; ambient conditions, design, and social factors for PEQ; and waiting time, tangibles, valence, and (empirically) social factors for outcome. Each of the 9 subdimensions is rated on reliability (R), responsiveness (RS), and empathy (E).	You can count on the employees at XYZ being friendly (R item of the attitude subdimension of IQ), XYZ understands that its atmosphere is important to me (E item of the ambient conditions subdimension of PEQ), XYZ tries to keep my waiting time to a minimum (RS item of the waiting time subdimension of OQ). All items (27 in total) rated on 7-point strongly agree-disagree scales. Global perceptual measures (2 each) for overall service quality and the primary dimensions were	Third-order factor model in which service quality is formed by IQ, PEQ, and OQ, which in turn are reflected in three subdimensions each (one with a double loading). Each of the 9 subdimensions is reflectively indicated by an R, RS and E item. Alternative models were considered as well.	Quantification of a firm, or components of a firm such as employees (object), on service quality (attribute) based on customer ratings (rater).

			collected as well. No reversed items.		
Salesperson attributions and behavioral intentions of successful and unsuccessful sales calls (Dixon, Spiro, and Jamil, 2001)	No explicit definition, but the goal is to develop a complete set of attributional and behavioral scales for sales success and failure.	Five dimensions of attributions: effort, ability, task, strategy, and luck. Five dimensions of behavioral intentions: no change, increase effort, change strategy, seek assistance, and avoid situation.	Three items each for the five attribution dimensions (e.g., I put forth the effort needed to make this sale, I have the necessary skills, I picked the strategy for this type of client) and the five behavioral intention dimensions (e.g., I would do the same thing, I will work much harder, I will avoid such situations in the future), rated on 6-point strongly agree-disagree scales. Different items for successful and unsuccessful sales experiences. No reversed items.	Reflective correlated 10-factor model (separately for successful and unsuccessful sales experiences). In addition, structural model in which the attribution dimensions predict the behavioral intention dimensions.	Quantification of a salesperson (object) on attributions and behavioral intentions [for successful and failed sales calls] (attribute) based on self-report (rater).
Global marketing strategy (GMS) (Zou and Cavusgil, 2002)	GMS as "the degree to which a firm globalizes its marketing behaviors in various countries through standardization of the marketing-mix variables, concentration and coordination of marketing activities, and integration of competitive moves across the markets" (p. 43).	Eight dimensions: product standardization, promotion standardization, standardized channel structure, standardized price (eventually dropped), concentration of marketing activities, coordination of marketing activities, global market participation, and integration of competitive moves.	Total of 20 items (ranging from 1 to 4 items per dimension), rated on 7-point strongly agree-disagree scales or 7-point bipolar scales. Example items are Main features of our product are standardized across major markets in the world, We develop similar channel structure for distributing any product in different country markets, and After-sale services (not coordinated at all to highly coordinated). Both items for one dimension (promotion standardization) are reversed.	Fully reflective second-order factor model.	Quantification of a firm or BU (object) on GMS (attribute) based on key-informant reports (rater).
Centrality of visual product aesthetics	CVPA as "the overall level of significance	Initially, four dimensions were hypothesized, but then three	Owning products that have superior designs makes me	A reflective correlated three-	Quantification of a consumer

(CVPA) (Bloch, Brunel, and Arnold, 2003)	that visual aesthetics hold for a particular consumer in his/her relationships with products" (p. 552).	dimensions were suggested: value (personal and social value of design), acumen (aesthetic sensibility), and response (level of response to design aesthetics). Ultimately, a one-factor model was retained.	feel good about myself (value, 4 items), Being able to see subtle differences in product designs is one skill I have developed over time (acumen, 4 items), and Sometimes the way a product looks seems to reach out and grab me (response, 3 items), rated on 5-point Likert scales. No reversed items.	factor model and a fully reflective second-order factor model were estimated, but because of high factor correlations, ultimately a one-factor model was proposed.	(object) on CVPA (attribute) through self-report (rater).
Need for touch (NfT) (Peck and Childers, 2003)	NFT as "a preference for the extraction and utilization of information obtained through the haptic system" (p. 431)	Two dimensions: instrumental touch (prepurchase touch reflecting outcome-directed touch with a salient purchase goal) and autotelic touch (touch as an end in and of itself).	I place more trust in products that can be touched before purchase (instrumental, 6 items) and When walking through stores, I can't help touching all kinds of products (autotelic, 6 items), rated on 7-point strongly agree/disagree scales. No reversed items.	Reflective correlated 2-factor model.	Quantification of an individual (object) on NFT (attribute) through self-report (rater).
Hedonic and utilitarian dimensions of attitude (Voss, Spangenberg, and Grohmann, 2003)	No definition provided.	Two dimensions: hedonic dimension "resulting from sensations derived from the experience of using products", utilitarian dimension "derived from functions performed by products" (p. 310).	Effective/ineffective or functional/not functional for utilitarian (5 items) and fun/not fun or delightful/not delightful for hedonic (5 items), rated on semantic differential scales. Possibly one reversed item (in terms of direction of response scale).	Reflective correlated 2-factor model, but also reflective second-order factor model (presumably with fixed second-order loadings).	Quantification of a product or brand (object) on hedonic or utilitarian dimension of attitude (attribute) through consumer ratings (rater).
Connectedness (with TV programming) (C) (Russell, Norman, and Heckler, 2004)	Connectedness as the "level of intensity of the relationship(s) that a viewer develops with the characters and contextual settings of a program in the para-social television environment" (p. 152.	Six dimensions: aspiration (A), modeling (M), imitation (I), fashion (F), paraphernalia (P), and escape (E).	I would love to be an actor in ___ (A, 2 items), I learn how to handle real life situations by watching ___ (M, 3 items), I find myself saying phrases from ___ when I interact with other people (I, 3 items), I like the clothes they wear on ___ (F, 3 items), I read books if they are related to ___ (P, 2	Fully reflective factor model with six first-order and 1 second-order factor (although this model fits significantly worse than a correlated six-factor model).	Quantification of a consumer (object) on Connectedness (attribute) via self-report (rater).

			items), and Watching ___ is an escape for me (E, 3 items), rated on 5-point strongly agree-disagree scales. No reversed items.		
Customer relationship management process (CRMP) (Reinartz, Krafft, and Hoyer, 2004)	CRMP as "the systematic and proactive management of relationships as they move from beginning (initiation) to end (termination), with execution across the various customer-facing contact channels" (p. 295).	Three primary dimensions (relationship initiation or RI, maintenance or RM, and termination or RT), each composed of 3, 4, and 2 subdimensions, respectively).	We have a formal system for identifying potential customers (7, 4, and 4 items for the subdimensions of RI), We have a formal system for determining which of our current customers are of the highest value (4, 7, 5, and 4 items for RM), and We have a formal system for identifying nonprofitable or lower-value customers (1 and 3 items for RT), rated on 7-point Likert scales. No reversed items.	The three primary dimensions are formatively measured by the 9 subdimensions, and the 9 subdimensions are in turn formatively measured by the observed items.	Quantification of an SBU's (object) CRMP (attribute) through key informant reports (rater).
Tightwads and spendthrifts (ST-TW) (Rick, Cryder, and Loewenstein, 2008)	ST-TW refers to "individual differences in the tendency to experience a pain of paying" (p. 769).	Conceptualized as a bipolar dimension varying from tightwaddism at the low end to spendthriftiness at the high end.	Which of the following descriptions fits you better: 1=tightwad (difficulty spending money) to 11=spendthrift (difficulty controlling spending) plus three other items asking people to rate themselves in terms of fit and similarity to hypothetical descriptions. Two reversed items.	Reflective one-factor model.	Quantification of an individual (object) on ST-TW (attribute) through self-report (rater).
Elaboration on potential outcomes (EPO) (Nenkov, Inman, and Hulland, 2008)	EPO is a "generalized predisposition toward thinking about consequences" (p. 126).	Initially, four dimensions were hypothesized, but empirically three dimensions emerged: generation of potential consequences and evaluation of their importance and likelihood (GPC); encoding consequences with a positive outcome focus (POF);	I always try to assess how important the potential consequences of my decisions might be (GPC, 6 items), I keep a positive attitude that things always turn out all right (POF, 3 items), and When thinking over my decisions I focus more on their negative	Reflective correlated 3-factor model.	Quantification of an individual (object) on EPO (attribute) through self-report (rater).

		encoding consequences with a negative outcome focus (NOF).	end results (NOF, 4 items), rated on 7- or 5-point strongly agree/disagree scales. No reversed items.		
Consumer emotional intelligence (CEI) (Kidwell, Hardesty, and Childers, 2008)	CEI as a "person's ability to skillfully use emotional information to achieve a desired consumer outcome" (p. 154).	Four dimensions: perceiving (P), facilitating (F), understanding (U), and managing (M) emotions.	Respondents are asked to indicate the amount of specific emotions expressed in pictures of products (P, 5 items), the usefulness of experiencing certain emotions in particular situations (F, 4 items), to pick a specific emotion appropriate for a particular situation (U, 5 items), or to judge the effectiveness of a certain behavior in a particular situation (M, 4 items) (see www.ceis-research.com for the 18-item scale and scoring instructions).	Fully reflective second-order factor model.	Quantification of a consumer (object) on emotional intelligence (attribute) based on self-report (rater).
Compulsive buying (Ridgway, Kukar-Kinney, and Monroe, 2008)	A "consumer's tendency to be preoccupied with buying that is revealed through repetitive buying and a lack of impulse control over buying" (p. 622).	Two dimensions: obsession with buying leading to repetitive buying (obsessive-compulsive disorder) and lack of control over the urge to buy (impulse-control disorder)	Others might consider me a 'shopaholic' (obsessive-compulsive buying, 3 items) and I buy things I did not plan to buy (impulse buying, 3 items), rated on 7-point strongly agree-disagree or never-very often scales. No reversed items.	Reflective correlated 2-factor model.	Quantification of a consumer (object) on compulsive buying (attribute) through self-report (rater).
Brand engagement in self-concept (BESC) (Sprott, Czellar, and Spangenberg, 2009)	BESC as consumers' "propensity to include important brands in their self-concept" (p. 92).	Unidimensional conceptualization.	I consider my favorite brands to be part of myself (8 items in total), measured on 7-point strongly agree-disagree scales. No reversed items.	Reflective one-factor model.	Quantification of a consumer (object) on BESC (attribute) through self-report (rater).
Gender dimensions of brand personality (GDBP) (Grohmann, 2009)	GDBP as "the set of human personality traits associated with masculinity and	Two dimensions: masculine brand personality (MBP) and feminine brand personality (FBP).	Example items include brave or aggressive for MPB (6 items) and sensitive or tender for FBP (6 items), rated on 9-	Reflective correlated two-factor model.	Quantification of brands (object) on GDBP (attribute)

	femininity applicable and relevant to brands" (p. 106).		point not at all-extremely descriptive scales. No reversed items.		by consumers (rater).
Propensity to plan (PTP) (Lynch *et al*, 2009)	PTP as "individual differences in (a) frequency of forming planning goals, (b) frequency and depth of thinking through means of implementing subgoals, (c) use of activities and props to serve as reminders and to help see the big picture and constraints, and (d) personal preference to plan" (p. 109).	Unidimensional conceptualization.	I set financial goals for the next few days (1-2 months) for what I want to achieve with my money (time), rated on 6-point strongly agree-disagree scales. The total scale consists of 6 items, and there are separate versions for planning for money and time in either the short or long run (4 different versions). No reversed items.	Reflective one-factor model.	Quantification of a consumer (object) on PTP (attribute) by self-report (rater).
Material possession love (MPL) (Lastovicka and Sirianni, 2011)	MPL as a "property of a consumer's relationship with a specific psychologically appropriated possession, reflecting the nature and degree of a consumer's positive emotional attachment to an object" (p. 324).	Three dimensions: passion (P), intimacy (I), and commitment (C). The three components define seven forms of love (either singly or in various combinations).	Just thinking about (my car) "turns me on" (P, 6 items), I enjoy spending time on (my car) (I, 8 items), I would like to always keep (my car) (C, 3 items), rated on 6-point definitely agree-disagree scales. No reversed items.	Reflective correlated three-factor model.	Quantification of a consumer (object) on MPL (attribute) by self-report (rater).
Brand love (Batra, Ahuvia, and Bagozzi, 2012)	No explicit definition provided, but brand love is conceptualized as a consumer-brand relationship that corresponds to a brand love prototype	Ten major components (based on qualitative research): high quality (eventually treated as an antecedent), linkages to strongly held values, beliefs that the brand provided intrinsic rather than extrinsic rewards, use of the loved	14 factors are eventually distinguished measured by 57 items, presumably rated on not at all to very much and other scales. Three higher-order factors (self-brand integration, passion-driven behaviors, positive emotional connection)	Fully reflective third-order factor model in which brand love is reflected in three second-order and four first-order factors.	Quantification of a consumer or brand (object) on brand love (attribute) by self-report (rater).

	consisting of 10 components.	brand to express both current and desired self-identity, positive affect, a sense of rightness and a feeling of passion, an emotional bond, investments of time and money, frequent thought and use, and length of use.	measured by three subfactors each (e.g., says something about who you are, feel myself craving to use it, feels like old friend), and five first-order factors (long-term relationship, anticipated separation distress, attitude valence, and two attitude strength factors, e.g., will be using for a long time, like-dislike, feel lots of affection toward it). No reversed items.		
Brand schematicity (BS) (Puligadda, Ross, and Grewal, 2012)	BS as a "consumer predisposition to process information using brand schema" (p. 115).	Unidimensional conceptualization.	When I am considering products, the brand name is more important to me than any other information or I like to surround myself with recognizable brand names at home, presumably rated on 9-point completely agree-disagree scales. Four of 10 items are reversed.	Reflective one-factor model.	Quantification of a consumer (object) on BS (attribute) through self-report (rater).
Lay rationalism (LR) (Hsee *et al*., 2015)	LR as "using reason rather than feelings to guide decisions" (p. 134).	Unidimensional conceptualization.	When making decisions, I think about what I want to achieve rather than how I feel, rated on 6-point strongly agree-disagree scales. Two of 6 items are reversed.	Reflective one-factor model.	Quantification of a consumer (object) on lay rationalism (attribute) through self-report (rater).
New product design (Homburg, Schwemmle, and Kuehnl, 2015).	Product design as "a set of constitutive elements of a product that consumers perceive and organize as a multidimensional construct comprising the three dimensions of aesthetics,	Three dimensions: aesthetics (A), functionality (F), and symbolism (S).	The product is good looking (A, 3 items), The product seems to be capable of doing its job (F, 3 items), and The product would help me in establishing a distinctive image (S, 3 items), rated on 5-point strongly agree-disagree scales. No reversed items.	Reflective correlated 3-factor model.	Quantification of a product (object) on product design (attribute) by consumers (rater).

	functionality, and symbolism" (p. 44).				
Customer inspiration (CI) (Böttger *et al*. 2017)	CI as "a customer's temporary motivational state that facilitates the transition from the reception of a marketing-induced idea to the intrinsic pursuit of a consumption-related goal" (p. 117).	Two dimensions: epistemic activation component ("inspired by") and intention component ("inspired to").	My imagination was stimulated (inspired-by, 5 items); I was inspired to by something (inspired-to, 5 items), rated on 7-point strongly agree-disagree scales. No reversed items.	Reflective correlated 2-factor model (although conceptually inspired-by is treated as an antecedent of inspired-to).	Quantification of a customer (object) on CI (attribute) through self-report (rater).
Locavorism (Reich, Beck, and Price, 2018)	Locavorism as the preference for local foods.	Three dimensions: lionization of local foods (L), opposition to long-distance food systems (O), and communalization of food economies (C).	Locally produced foods just taste better (L, 3 items), Large, global food systems are destined to fail (O, 4 items), I like to support local farmers whenever possible (C, 4 items). One L item is reversed.	Reflective correlated 3-factor model.	Mainly quantification of a consumer (object) on locavorism (attribute) by the consumer (rater), but for some items the object is foods, producers or food systems, and the attributes are quality or societal outcomes.

Table 3.1: Classification of methods used to measure satisficing

	DEDICATED MEASURES Special items or scales are included in the questionnaire to measure satisficing	NO DEDICATED MEASURES Satisficing is inferred from respondents' answers to substantive questions
DIRECT MEASUREMENT Satisficing is assessed directly by measuring respondents' tendency to minimize time and effort when answering questions	CATEGORY 1 Self-reported effort (e.g., I carefully read every survey item).	CATEGORY 2 Response time Gaze duration (eye-tracking)
INDIRECT MEASUREMENT Satisficing is assessed indirectly based on the presumed consequences of respondents' attempts to minimize time and effort on the quality of responses	CATEGORY 3 Quality of responses to special items or scales (e.g. bogus items, instructed response items)	CATEGORY 4 Quality of responses to substantive questions (e.g., outlier analysis, lack of consistency of responses, excessive consistency of responses)

Table 3.2: Some key recommendations on item wording

RECOMMENDATION	EXAMPLE
Replace unfamiliar words (esp. low-frequency words) with more familiar words	'Physical pain' is better than 'somatic pain'
Replace vague or imprecise relative terms with more precise terms	'The last four weeks' is better than 'recently'
Replace vague or ambiguous noun phrases with more specific terms	'Did you go to the theater?' is better than 'Did you attend cultural events?'
Simplify syntax and avoid complex logical structures	'Before I do something, I consider all possible outcomes' is better than 'Before I act, I consider what I will gain or lose in the future as a result of my actions.'
Avoid low syntactic redundancy	'Unions are important to secure the jobs of employees' is easier to understand than 'Unions are important for the job security of employees'.

Note: Based on Graesser *et al*. (2006); Graesser *et al*. (2000); Hardy and Ford (2014); and Lenzner (2012, 2014).

Table 3.3 An overview of common response styles

Response style	*Definition and synonyms*	*Theoretical explanations*	*Measurement*
Acquiescence response style (ARS)	The tendency to *agree* with items regardless of content. Also called agreement tendency, yeasaying, or positivity bias.	▪ Characteristic of stimulation-seeking extraverts who have a tendency to impulsively accept statements. ▪ Due to uncritical endorsement of statements by respondents who are low in cognitive abilities or have low status. ▪ More common for items that are ambiguous, vague, or neutral in desirability, or for issues about which respondents are uncertain. ▪ Most likely when respondents lack adequate cognitive resources because of distraction, time pressure, etc.	*Two general approaches*: ▪ Extent of agreement with many items that are heterogeneous in content. ▪ Extent of agreement with both regular-keyed and reversed-keyed items within the same substantive scale (before reversed items have been recoded).
Disacquiescence response style (DARS)	The tendency to *disagree* with items regardless of content. Also called disagreement tendency, naysaying, or negativity bias.	▪ Characteristic of controlled and reflective introverts trying to avoid external stimulation.	Same as acquiescence, except that disagreement is assessed instead of agreement.
Net acquiescence response style (NARS)	The tendency to show greater acquiescence than disacquiescence. Also called directional bias.	See explanations for acquiescence and disacquiescence.	In general, acquiescence minus disacquiescence. Most commonly measured as the mean response across many heterogeneous items.

Extreme response style (ERS)	The tendency to endorse the most extreme response categories regardless of content.	▪ Reflection of rigidity, intolerance of ambiguity, and dogmatism. ▪ Associated with higher levels of anxiety and possibly deviant behavior. ▪ Characteristic of respondents with less differentiated cognitive structures and poorly developed schemas. ▪ Greater for "meaningful" stimuli (i.e., stimuli that are important or involving to respondents). ▪ Greater for self-relevant items, especially among respondents with an independent self-construal (vs. interdependent self-construal).[a]	Number or proportion of heterogeneous items on which the respondent endorses the most extreme (positive or negative) scale categories. Greenleaf (1992b) suggests that the items should be uncorrelated and have equal extreme response proportions. In addition, the mean response to an item should be close to the midpoint of the scale.
Midpoint responding (MPR)	The tendency to use the middle scale category regardless of content.	▪ Due to evasiveness (desire not to reveal one's true opinion), indecision (uncertainty about one's position), or indifference (disinterest in an issue).	Number or proportion of heterogeneous items on which the respondent endorses the middle scale category.

Note: Based on Baumgartner and Steenkamp (2001); see the original article for detailed references. [a] is based on Cabooter *et al*. (2016).

Table 4.1 A classification of observed variables based on continuity and level of measurement with illustrative examples

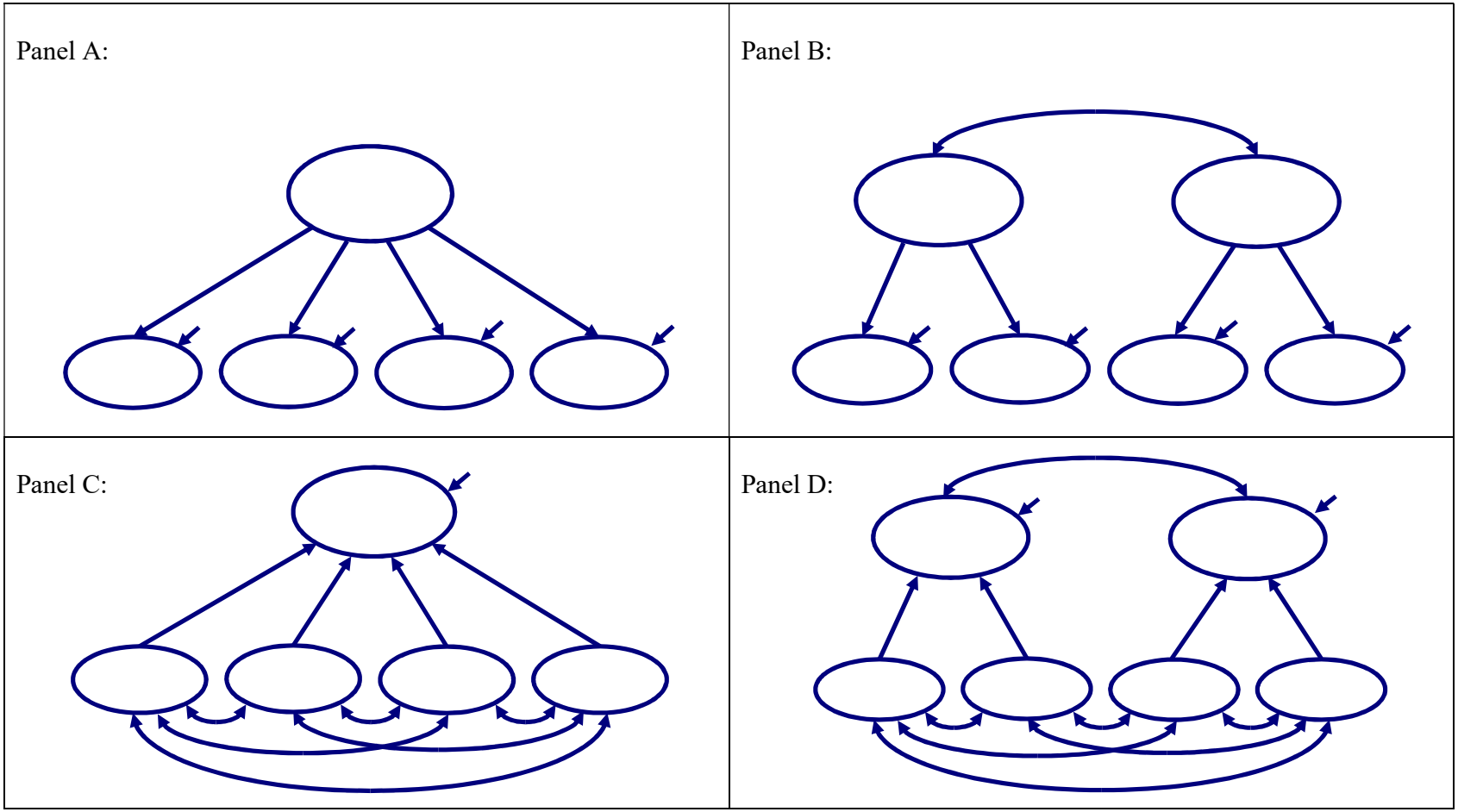
	Nominal scales	Ordinal scales	Metric scales
Discrete variables	Gender identity measured as 1 = male, 2 = female, 3 = transgender, and 4 = do not identify with a particular gender.	Extent of (dis)agreement measured on a 5-point scale (e.g., 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree).	Number of coupons redeemed during the last trip to the supermarket.
Continuous variables	n.a.	Degree of liking measured on a 0 to 100 slider scale calibrated in millimeters.	Response time measured to the nearest millisecond.

Table 4.2 Measurement analysis for the 18-item MVS scale

	DF	ML χ^2 value	SRMR	RMSEA	CFI	TLI	BIC
Exploratory factor model with three factors	102	239.99	.038	.049	.940	.910	26650.57
Congeneric confirmatory factor model with 3 trait factors	132	415.24	.052	.062	.876	.856	26636.30
Models with method factors for the regular and/or reversed items							
3 trait factors, 1 method factor	114	228.19	.035	.043	.950	.933	26562.96
3 trait factors, 1 method factor (equal loadings)	131	251.94	.040	.041	.947	.938	26479.32
3 trait factors, 1 method factor for regular items	123	259.92	.042	.045	.940	.926	26537.84
3 trait factors, 1 method factor for regular items (equal loadings)	131	274.91	.046	.045	.937	.927	26502.29
3 trait factors, 1 method factor for reversed items	123	239.93	.038	.041	.949	.936	26517.85
3 trait factors, 1 method factor for reversed items (equal loadings)	131	262.73	.043	.043	.942	.933	26490.11
3 trait factors, 2 uncorrelated method factors for regular and reversed items	114	202.03	.035	.037	.962	.948	26536.80
3 trait factors, 2 uncorrelated method factors for regular and reversed items (equal loadings)	130	253.41	.041	.041	.946	.937	26487.11
3 trait factors, 2 correlated method factors for regular and reversed items	113	196.14	.033	.036	.964	.951	26537.23
3 trait factors, 2 correlated method factors for regular and reversed items (equal loadings)[1]	129	251.83	.040	.041	.946	.936	26491.85
3 trait factors, 2 implicit ARS method factors	113	223.22	.035	.042	.952	.935	26564.31
3 trait factors, 2 implicit ARS method factors (equal loadings)	129	250.04	.040	.041	.947	.937	26490.05
Models with correlated uniquenesses for the regular and/or reversed items							
3 trait factors, correlated uniquenesses for regular items	96	181.44	.036	.040	.963	.940	26629.92
3 trait factors, correlated uniquenesses for reversed items	96	195.25	.034	.043	.957	.931	26643.73
3 trait factors, separate correlated uniquenesses for regular and reversed items[1]	60	110.55	.027	.039	.978	.944	26786.45

Note: DF = degrees of freedom; ML = maximum likelihood; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker Lewis index; BIC = Bayesian information criterion. Models marked with the superscript 1 contain improper estimates.

Figure 2.1 Illustrative higher-order factor specifications



Note: Illustrative examples of higher-order factor structures underlying four first-order factors. The models are not measurement models, since no observed variables are shown, but higher-order factor models. The models in panels A and C are unidimensional second-order factor models, the models in panels B and D correlated two-factor second-order factor models. The models in panels A and B are reflective factor models, the models in panels C and D formative factor models. The formative factor models are not identified as shown (see the discussion in section 4).

Figure 4.1 A congeneric measurement model

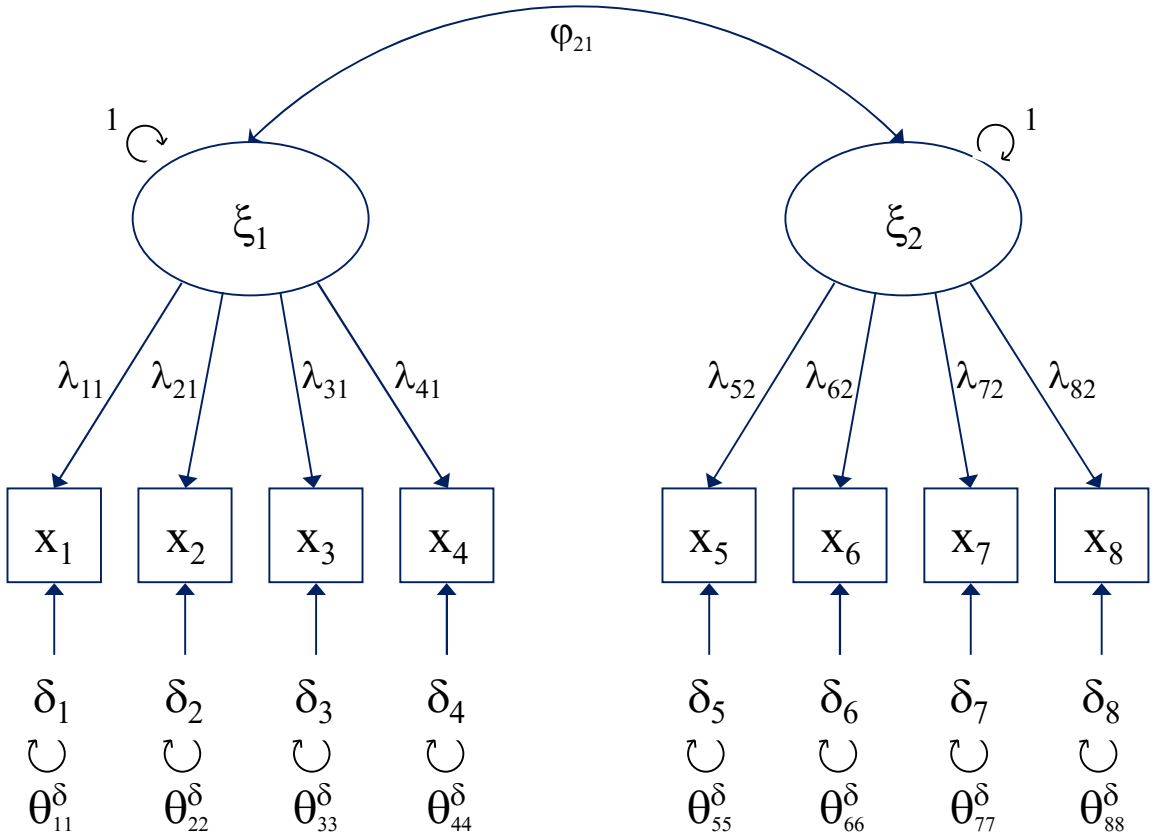


Figure 4.2 A classification of method effect models

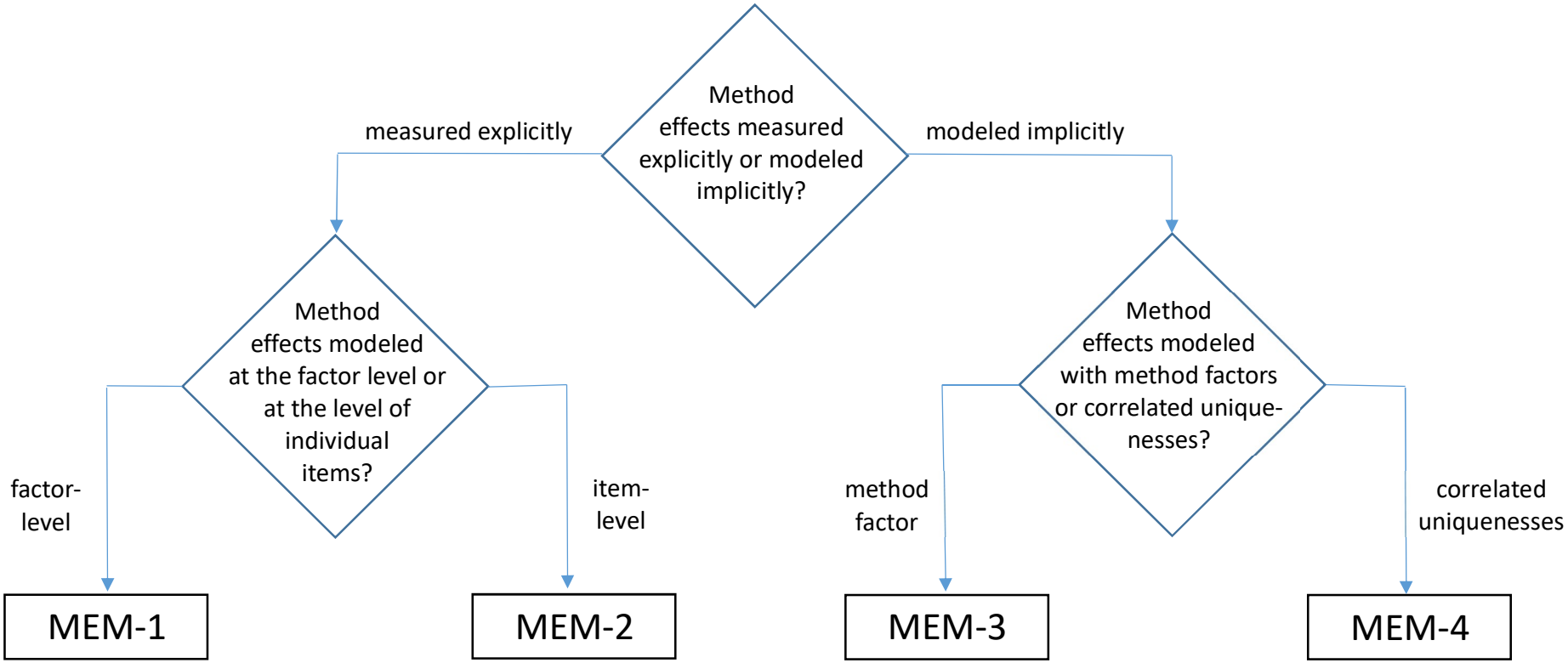
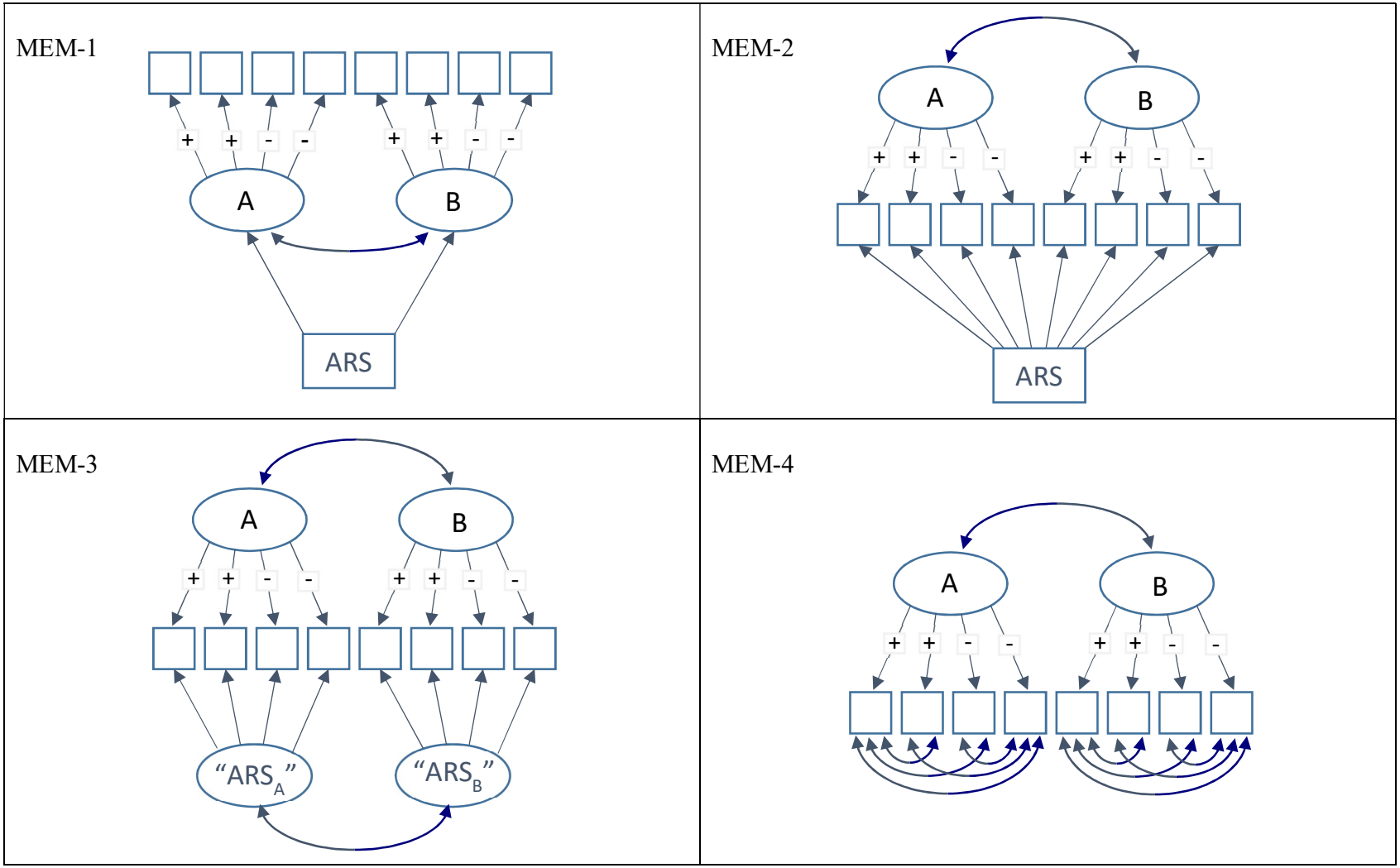


Figure 4.3 Illustrative examples of four types of method effect models



Note: A and B are two hypothetical constructs measured by four indicators each, two of which are reversed items. Since it is assumed that reversed items have not been recoded, regular items should have a positive loadings and reversed items should have a negative loading on the underlying substantive construct (as shown in the Figure). ARS refers to direct measure of acquiescence response style, "ARS_A" and "ARS_B" to separate inferred ARS factors for constructs A and B.

Figure 4.4 An illustrative multitrait multimethod model

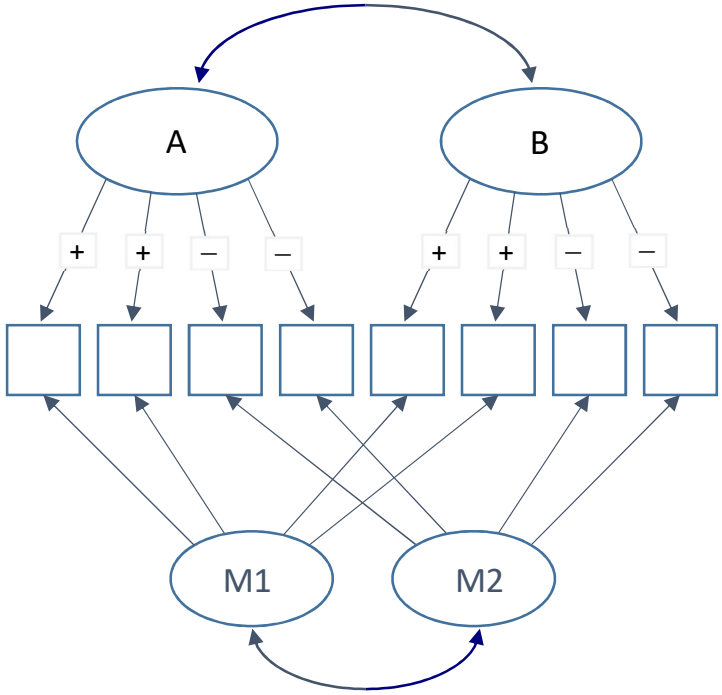
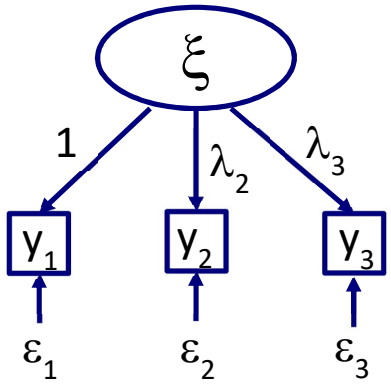


Figure 4.5 Illustrative examples of reflective and formative measurement models

Panel A: A reflective measurement model



Panel B: A formative measurement model