Chapter 10

Methodological issues in cross-cultural research

Hans Baumgartner & Bert Weijters

Hans Baumgartner, Smeal College of Business, The Pennsylvania State University, University

Park, PA 16802, USA, E-mail: jxb14@psu.edu.

Bert Weijters, Department of Personnel Management, Work and Organizational Psychology,

Ghent University, Dunantlaan 2, 9000 Ghent, Belgium, E-mail: bert.weijters@ugent.be.

Abstract

Regardless of whether the research goal is to establish cultural universals or to identify and explain cross-cultural differences, researchers need measures that are comparable across different cultures when conducting cross-cultural studies. In this chapter we describe two major strategies for enhancing cross-cultural comparability. First, we discuss *a priori* methods to ensure the comparability of data in cross-cultural surveys. In particular, we review findings on cross-cultural differences based on the psychology of survey response and provide suggestions on how to deal with these cultural differences in the survey design stage. Second, we discuss *post hoc* methods to ascertain data comparability and enable comparisons in the presence of threats to equivalence.

Keywords: Survey research, cross-cultural differences, questionnaire translation, response styles and response sets, measurement equivalence, corrections for lack of measurement invariance.

10.1 Introduction

With markets becoming increasingly globalized, cultural issues are taking a more central place in consumer psychology and business research. Two key motivations lead researchers to engage in cross-cultural research. First, researchers tackle questions of generalizability, aiming to find out whether theories and models initially developed and validated in one culture (typically the US) hold in other cultures as well (Dawar & Parker, 1994). Researchers often hope to establish 'strong theories' or 'universals' that are generally valid and are not limited to a specific (cultural) context (Dawar & Parker, 1994; Laczniak, 2015). Second, in case models and theories are found not to be universal, the research focus shifts to questions of differentiation, aimed at identifying and explaining differences in multivariate relations of interest across different groups of consumers, where the grouping is oftentimes defined by national culture (Steenkamp, 2001). Whether the research goal is to establish universals or to identify and explain cross-cultural differences, researchers need cross-culturally equivalent measures to realize their research goals. Making measurements comparable across different cultures has proven to be very challenging, however. In this chapter, key methodological challenges in cross-cultural research will be discussed, with a focus on issues related to threats to the cross-cultural comparability of survey data and possible solutions in terms of survey design and data analysis. Our emphasis will not be on sampling, data collection, and survey administration issues (see Usunier, van Herk, & Lee, 2017). Rather, we focus on the rapidly evolving literature studying differences in the way people interpret and respond to questions, the biases that can result from these differences, and the procedures researchers can use to prevent or control for these biases. We will distinguish two major strategies for enhancing cross-cultural comparability. First, we describe *a priori* methods

to ensure the comparability of data in cross-cultural surveys, especially in terms of survey design (including translation). In particular, we will discuss cross-cultural differences in the psychology of survey response, the consequences of these differences, and ways of dealing with them in the survey design stage. Second, we discuss *post hoc* methods to ascertain data comparability and enable comparisons in the presence of threats to equivalence, focusing on data-analytic issues.

10.2 Cross-cultural psychology of survey response and *a priori* survey design recommendations

To provide structure to our discussion, we distinguish between five key cognitive processes that are part of responding to survey questions (Tourangeau & Rasinski, 1988; Tourangeau, Rips, & Rasinski, 2000) and that have been found to be culture-specific at least to some extent (Schwarz, Oyserman, & Peytcheva, 2010): (1) comprehension (how people interpret the questions and specific concepts within them); (2) information retrieval (what information respondents recall from their memories in response to survey questions); (3) judgment (how respondents aggregate or summarize the information they retrieve); (4) response mapping (how participants map an internal judgment onto the response options that are provided); and (5) response editing (how participants alter their response in order to project a favorable image). These processes need not occur in a fixed sequence and may partially overlap. Also, even though we discuss specific cultural biases under each process, this mapping is far from perfect. Nevertheless, the model provides structure to the discussion and helps to organize a somewhat fragmented literature. In particular, under (1) comprehension, we discuss how subtle differences in the translation of items may lead to non-equivalence; under (2) retrieval, we explain culture-specific question context effects, where respondents' interpretation of what information is relevant for a given question is

partly driven by other questions in the survey; under (3) judgment, we point out that responses to reversed items tend to be differentially problematic for respondents from certain cultural backgrounds; under (4) response mapping, we describe differences in response styles related to culture and language; and under (5) response editing, we look at cultural differences in socially desirable responding.

10.2.1 Comprehension: Item translation

A key aspect of comparability is coming up with translations that make instructions and survey items equivalent in meaning across cultures. Typical questionnaire items consist of two parts: the stem of the item presenting the statement or question to which the respondent is asked to react, and the response scale used for recording the answers. Translation issues occur for both parts. We will focus on question design here and return to the issue of response scale translation in the section on response mapping.

Literal translations of a word sometimes do not relate to exactly the same concept across languages. Davidov, Meuleman, Cieciuch, Schmidt, and Billiet (2014) report some examples, including the case where tolerance toward immigrants involved in crime seemed to be much higher in Denmark than in other European countries. This ran counter to prior expectations based on electoral data, and it turned out to be due to an idiosyncratic translation of the word crime (which had a much broader meaning in Danish, because it included mild offenses such as violations of traffic rules). Weijters, Puntoni, and Baumgartner (2017) also give the example that a commonly used verb such as '(to) like' may not have equivalent counterparts in some other languages, including the French alternative 'aime(r)' (which could mean to 'like' or 'love', thus creating ambiguity). To partially circumvent such problems, Weijters, Puntoni, et al. (2017)

propose the principle that formulating key concepts in several distinct ways makes it possible to triangulate cross-linguistic variations in meaning. If multiple items related to the same construct all use different terms to refer to the same concept, measurement invariance testing (discussed in detail later in this chapter) can help identify non-equivalent translations. Using at least three linguistically distinct measures of the same construct is desirable, because group-differences can be triangulated (T. W. Smith, 2004; Smith, Mohler, Harkness, & Onodera, 2005; Weijters, Puntoni, et al., 2017).

A common approach for obtaining equivalent questionnaires across languages is the translation/back-translation procedure (Brislin, 1970). The method entails the following steps: (1) design a questionnaire in a source language; (2) translate it to one or multiple target languages by bilingual native speakers of the target languages; (3) translate the result back to the source language by bilingual speakers of the source language; and (4) resolve incidental differences based on a comparison of the initial and the back-translated questionnaire. Back-translation is helpful in identifying translation issues, but it cannot guarantee meaning equivalence across languages (Davidov et al., 2014; Douglas & Craig, 2007; Okazaki & Mueller, 2007). It has been suggested that a more collaborative approach may be preferable (Janet A. Harkness et al., 2010; Janet A Harkness, Edwards, Hansen, Miller, & Villar, 2010; Janet A. Harkness, Pennell, & Schoua-Glusberg, 2004). Douglas and Craig (2007) propose the following steps: (1) a committee tries to come up with equivalent key concepts to be used in the questionnaire; (2) two independent translators translate the questionnaire into the target language; (3) the translated questionnaire is pre-tested; and (4) the translation and pre-testing steps are repeated until equivalence is realized. This means that translations are assessed in terms of comprehension, clarity and coverage. Qualitative pretests typically yield richer insights into

translation issues, but quantitative pretests can help identify potentially problematic items in terms of internal consistency and/or factor structure (including measurement invariance, as discussed later). Relative to back-translation, this iterative collaborative procedure will typically require more resources. Also, it demands participation from a team of researchers who have extensive experience with questionnaire design and the languages involved. For example, in the so-called TRAPD team translation approach (TRAPD stands for Translation, Review, Adjudication, Pretesting, and Documentation) translators, reviewers, and adjudicators work together to produce a target instrument from a source instrument (Harkness, Pennell, & Schoua-Glusberg, 2004).

A guiding principle in designing cross-culturally valid questionnaires is *decentering*. Decentering is defined as the simultaneous development of the same instrument in several languages and/or cultures from the initiation of the project. This requires researchers to transcend their reference frame (including the idea of a source language), as opposed to practicing what has been called 'research imperialism or safari research' (T. W. Smith, 2004). A decentered approach also demands the cooperation of researchers who have a background in each of the languages (T. W. Smith, 2004; Van de Vijver & Leung, 1997). Although decentering has some obvious advantages, it is difficult to implement when a survey has to be developed in many different languages, as in the European Social Survey (Harkness & Schoua-Glusberg, 1998). Some readily implementable suggestions related to item wording to facilitate translation are provided by Brislin (1986), including the following: avoid metaphors and colloquialisms; use short, simple sentences in active voice (rather than passive voice); repeat nouns instead of using pronouns; use specific rather than general and/or vague terms; and avoid complex sentence

structures, such as subordinate clauses and adverbs and prepositions specifying "where" or "when" (see also Van de Vijver & Hambleton, 1996).

10.2.2 Retrieval: Question context effects

Surveys constitute a form of information exchange situated in a social setting to which certain conversational norms apply (Schwarz, 1999; Schwarz et al., 2010). Most surveys contain multiple questions addressing related issues. For instance, a survey may probe satisfaction with different life domains such as academic achievement, marital satisfaction and overall life satisfaction. When answering such questions, respondents will try to figure out what information they are supposed to recall for each of them. They will try to provide information that is deemed relevant but not redundant with answers to other related questions (Schwarz, Strack, & Mai, 1991), in line with conversational norms (Grice, 1975).

When constructs are related in a part-whole sequence (e.g., when an item related to satisfaction with an aspect of life is followed by an item related to general life satisfaction), several mechanisms are at play. First, priming effects make the content retrieved in response to an item more accessible when retrieving content related to another item encountered later in the survey. This will result in higher consistency in responses (Salancik, 1984). Second, however, when questions are perceived as being related to each other, conversational norms of non-redundancy disallow repeatedly using and reporting the same information. Consequently, survey participants tend to interpret a general question as referring to aspects other than the ones covered by a preceding specific question (Schwarz et al., 1991). But if several specific questions precede the general question, the general one will be interpreted as requesting a summary judgment.

Context effects like these are primarily driven by respondents' motivation to adhere to conversational norms and meet the researcher's expectations. Norm adherence and the extent to which people are motivated by others' expectations are partly culturally driven. For example, it has been argued that people in collectivistic cultures tend to have a more interdependent self-construal, defining themselves mainly in terms of their place in social networks and in relation to others, whereas people in individualistic cultures tend to have more of an independent self-construal, defining themselves mostly as autonomous individuals (Gardner, Gabriel, & Lee, 1999; Markus & Kitayama, 1991; Oyserman & Lee, 2008). Since attentiveness to others is more likely to be a self-defining goal when the self is thought of as interdependent with others (vs. independent of others), respondents from collectivistic cultures are believed to be more attentive to others and consequently more sensitive to context effects (Schwarz et al., 2010). In a priming experiment, Haberstroh, Oyserman, Schwarz, Kühnen, and Ji (2002) demonstrated that people with an interdependent self-construal were more likely than independence-primed participants to take the recipient's knowledge into account by avoiding to provide redundant information in a self-administered questionnaire. They then replicated these findings with participants from relatively more individualistic (Germany) and collectivistic (China) cultures. In this study, respondents answered questions on academic satisfaction and life satisfaction. If the life satisfaction question preceded the academic satisfaction question, no redundancy issue occurred, and correlations in both samples were very similar (both close to .50). However, when the academic satisfaction preceded the life satisfaction question, the correlation was .78 among German respondents, but only .36 among Chinese respondents. Such cultural differences in context effects are worrisome because they can result in spurious cultural effects. Researchers who are not aware of the cultural specificity of question context effects may erroneously

conclude that culture moderates a relation of interest, while in truth all that was moderated was the relation between responses to survey questions (not the latent construct the researcher aims to represent).

Clearly, researchers need to be aware of the way respondents interpret questions and the specific information that respondents retrieve to answer them. To avoid unintended idiosyncratic context effects, we suggest that researchers use the following preventive approaches: (1) provide explicit instructions on what is and what is not relevant to the question(s) at hand; (2) experimentally manipulate item context and incorporate item context effects as moderating variables into the conceptualization of research and in empirical models; and/or (3) conduct cognitive interviews to investigate the interpretation of survey items in all the cultures involved in the research.

10.2.3 Judgment: Reversed item bias

Once respondents have retrieved information from their memories, they need to somehow make a selection of what is and what is not relevant to the question at hand, and formulate a summary judgment. Evidence suggests that this process too is subject to cultural influences. In particular, when reversed items are used, East-Asian respondents have been found to show a tendency to agree with both the original items and their reversals. In a study among over 800 adults from the United States, Singapore, Thailand, Japan, and Korea, Wong, Rindfleisch, and Burroughs (2003) show that the cross-cultural measurement equivalence and construct validity of a materialism scale is threatened by the inclusion of reversed items. They establish similar problems with other scales that contain reversed items in a survey among approximately 400 Americans and East Asians. Wong et al. attribute this tendency to East Asians' Confucian belief system, which

encourages dialectical thinking, compromise, and a tolerance of contradictory beliefs. In line with this, Hamamura, Heine, and Paulhus (2008) argue that compared to those of European heritage, individuals of East-Asian heritage tend to exhibit greater ambivalence in their responses.

In sum, in surveys that include both Western and East Asian respondents, a valid comparison of survey answers may be hampered by differences in the extent to which the two groups of respondents tend to agree with reversed items. This can result in incomparable factor structures and measurement models, and even spill over to estimated relations between constructs, thus resulting in spurious moderating effects of national culture. However, since the use of (non-negated) reversed items can encourage better coverage of the content domain of a construct, and since in the absence of reversed items it is impossible to distinguish between non-substantive and substantive agreement (Weijters & Baumgartner, 2012), simply omitting reversed items is not a meaningful solution. What to do then?

Wong et al. show that an interrogative scale format (e.g., "How do you feel about people who own expensive homes, cars, and clothing?", rated on a scale from do not admire to greatly admire) lessens the problems introduced by reverse-worded items and thus enhances the cross-cultural applicability of such scales. We therefore recommend that researchers who collect data from Western and East Asian respondents reformulate scales in an interrogative format and subject them to pretesting before starting the actual study. In addition, including a factor that captures inconsistent responding to reversed items can be helpful in correctly modeling the data (Weijters, Baumgartner, & Schillewaert, 2013), although more work is needed to validate cross-cultural versions of such models.

10.2.4 Response: Culture and response styles

Once respondents have formulated an internal judgment in response to a survey question, they need to map this judgment onto the response options that are provided to them. This is where response styles come in, defined as respondents' disproportionate use of certain response categories regardless of item content. Most notably, respondents may make disproportionate use of the response categories on the positive side (acquiescence response style, or ARS) or negative side (disacquiescence response style, or DARS), the extremes (extreme response style, or ERS), or the middle of the scale (midpoint response style, or MRS).

If survey data from different cultures are differentially contaminated by response styles, the comparability of the data is compromised (Baumgartner & Steenkamp, 2001). The reason is that the survey responses would vary across cultures even if there were no true differences in the latent constructs they intend to measure (or vice versa). In addition, differential response style bias can cause spurious differences in univariate distributions as well as multivariate relations, including factor structures, correlations and regression weights. Thus, researchers may make erroneous conclusions in cross-cultural comparisons.

Cross-cultural differences in response styles have received much more research attention than the other cross-cultural biases we have discussed so far. Below, we point out some key findings discovered in this extensive literature.

First, research has established quite consistent response style differences between ethnic subcultures even within a single country (Bachman & O'Malley, 1984; Hui & Triandis, 1989; Marín, Gamba, & Marìn, 1992). Together, these results stress the importance of taking into account cross-cultural differences even within countries, as response styles can vary as a function

of differences in language use (Bachman & O'Malley, 1984), levels of acculturation (Marín et al., 1992), as well as response strategies and preferences (Morren, Gelissen, & Vermunt, 2012). Cognitive interviews with respondents from varying cultural or ethnic backgrounds can help researchers identify and avoid potentially ambiguous questions (Morren et al., 2012).

Second, many studies have reported response style differences between countries, and this has resulted in a consensus that response styles show cross-cultural variation (Baumgartner & Steenkamp, 2001; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Johnson, Kulesa, Cho, & Shavitt, 2005; Van Herk, Poortinga, & Verhallen, 2004; Van Rosmalen, Van Herk, & Groenen, 2010). Some studies have compared response styles for two or a small number of countries (see Harzing, 2006, for an overview of several such studies). Others have tried to relate cross-national response style differences to other cross-national variables, including cultural dimensions, in multi-country studies (Chen, Lee, & Stevenson, 1995; De Jong et al., 2008; Harzing, 2006; Johnson et al., 2005; P. B. Smith, 2004; Van Herk et al., 2004). Based on an extensive review of this literature, Baumgartner and Weijters (2015) conclude that the findings on cross-cultural variation in response styles are most consistent for individualism-collectivism: Individualism is negatively associated with (N)ARS and MRS, and tends to have a positive relationship with ERS, whereas there is less consistency across studies for the other dimensions.

Third, experimental research can provide stronger evidence about, and deeper insights into, the mechanisms leading to cross-cultural response style differences. Cabooter, Millet, Weijters, and Pandelaere (2016) experimentally replicate the finding that people with an independent self-construal generally answer more extremely to survey items than those with an interdependent self-construal, and they demonstrate that this holds especially when the items are self-relevant and processed more fluently. Self-concept clarity drives the effect of self-construal

on extreme responding: People with an independent self-construal have a higher level of self-concept clarity, and greater clarity induces higher ERS.

Finally, it is important to bear in mind that in many cross-national studies on response style differences, language effects need to be entertained as potential alternative explanations. Language relates to response styles in two major ways. First, respondents vary in terms of the level of fluency with which they use a language. Harzing (2006) shows that responding to a survey in one's native language results in lower MRS and higher ERS than taking the survey in English (as a second language). Furthermore, second language fluency positively relates to ERS and negatively relates to MRS. Apparently, language competence makes respondents more willing to respond more extremely, possibly because they feel more confident. Interestingly, in surveys dealing with emotions, de Langhe, Puntoni, Fernandes, and van Osselaer (2011) demonstrate a tendency for respondents to report more intense emotions when evaluating consumption experiences and products on rating scales that are not expressed in their native language. This 'Anchor Contraction Effect' (or ACE) occurs because bilinguals perceive emotional scale anchors in their non-native language as less intense than the same emotional anchors in their native language. Because ratings are typically provided relative to these scale anchors, second language rating scales yield more extreme ratings. To circumvent this non-equivalence, de Langhe et al. (2011) offer some suggestions. The most appropriate solution is to make sure that all respondents answer questionnaires in their native language. But in two studies, de Langhe et al. (2011) also demonstrate the effectiveness of using emoticons or colors. Emoticons are recommended for measuring specific emotions, in particular basic emotions that can be easily portrayed with stylized facial expressions. Colors are most suitable for abstract or complex emotional concepts (such as pity or emotionality). Unfortunately, associations between

colors and emotions are partly culture-specific, so colors may be vulnerable to cross-cultural

differences in interpretation (Hupka, Zaleski, Otto, Reidl, & Tarabrina, 1997).

A second link between language and response styles is that language can affect responses

via non-equivalent response category labels (such as strongly disagree, disagree, etc.).

Researchers need to pay special attention to the issue of designing response scale formats that are

cross-culturally and cross-linguistically equivalent, because often a common response scale is

used throughout the questionnaire, which can introduce systematic between-group biases

(Weijters, Baumgartner, & Geuens, 2016; Weijters, Geuens, & Baumgartner, 2013). Consider a

situation in which a researcher uses the response category label 'strongly agree' in English and is

wondering how to translate this label into French. One option would be 'fortement d'accord',

which is a literal translation with presumably similar intensity, but which does not sound familiar

in French. Another option would be 'tout à fait d'accord' (literally 'completely agree'), which is

somewhat more intense, but which sounds much more familiar in French. In a series of studies,

Weijters, Geuens, et al. (2013) show that response categories with labels that do not sound

familiar in a given language will be endorsed less frequently. The intensity of the label is less

impactful, so if researchers need to trade off cross-linguistically equivalent familiarity with

equivalent intensity, familiarity overrules intensity. If unfamiliar labels are used for the response

scale endpoints in one language, this may suppress univariate and multivariate response

distributions, which can result in spurious cross-linguistic differences. Since language is often

confounded with culture, such differences can easily be misconstrued as evidence for substantive

cross-cultural variation. This is especially true since measurement invariance testing is often

unable to detect uniform bias (i.e., bias that is the same across multiple items, which is typically

the case with response scale effects) (Weijters, Baumgartner, et al., 2016).

10.2.5 Editing: Culture and SDR

When answering survey questions, respondents sometimes report what makes them look good rather than what is true. This phenomenon is referred to as Socially Desirable Responding (SDR) (Steenkamp, De Jong, & Baumgartner, 2010). Since, as pointed out by Baumgartner and Weijters (2015), perceptions of what looks good depend on prevailing cultural norms, it follows that cross-cultural differences in SDR are likely to exist.

SDR is not a one-dimensional construct. Probably the most important distinction is that between egoistic response tendencies, where respondents overestimate their own agency-related capabilities, versus moralistic response tendencies, where respondents project an image of themselves that optimally aligns with social norms. Steenkamp et al. (2010) show that these dimensions correspond, respectively, to the self-deceptive enhancement (SDE) and impression management (IM) dimensions in the Balanced Inventory of Desirable Responding, which was developed and validated by Paulhus (1991). Both aspects of SDR can unconsciously emerge in the absence of situational demands to project a positive image, but can additionally be strengthened in response to situational pressures such as high stakes contexts, public disclosure, or questions related to sensitive topics (Steenkamp et al., 2010).

SDR tendencies have been found to be associated with national culture, most importantly individualism/collectivism. Lalwani, Shavitt, and Johnson (2006) show that collectivism is positively related to impression management, whereas individualism is positively related to the self-enhancement component of SDR. Steenkamp et al. (2010) also find that IM is higher in collectivist countries than in individualist countries, but they find a negative relation between individualism and egoistic response tendencies. Uskul, Oyserman, and Schwarz (2010) make a

further distinction between Confucian-based collectivist cultures and honor-based collectivist cultures. The former emphasize fitting in and saving face, the latter emphasize positive presentation of self and one's in-group. Uskul, Oyserman, Schwarz, Lee, and Xu (2013) compare SDR among participants from a culture of modesty (China), a culture of honor (Turkey), and a culture of positivity (U.S.), who rated their own or someone else's success in life. The scale format was also manipulated to imply a continuum from failure to success (-5 to +5, i.e., bipolar scale format) or varying degrees of success (0 to 10, i.e., unipolar scale format). Response patterns depended on the interaction of culture and rating format. Uskul et al. (2013) conclude that "Americans, sensitive to the possibility of negativity, rated all targets more positively in the bipolar condition. Chinese were modesty-sensitive, ignoring the implications of the scale, unless rating strangers for whom modesty is irrelevant. Turks were honor-sensitive, rating themselves and their parents more positively in the bipolar scale condition and ignoring scale implications of rating strangers."

To conclude, we suggest some guidelines in order to minimize cross-cultural biases due to SDR. First, where this is possible, anonymity should be optimized and clearly communicated to respondents to reduce the motivation for SDR. Second, questions need to be pre-tested to assess their vulnerability to SDR and adapted if necessary. Third, where this is useful and feasible, questions can be worded indirectly, for instance by referring to a third person form (Fisher, 1993; Luchs, Naylor, Irwin, & Raghunathan, 2010). Finally, if a socially sensitive topic is the focus of the survey, researchers should consider using specifically designed measurement techniques such as the bogus pipeline (Roese & Jamieson, 1993), item randomized response (De Jong, Pieters, & Fox, 2010) or the Dual-Questioning-Technique Design (De Jong, Fox, & Steenkamp, 2015).

10.3 Post hoc methods to ascertain data comparability and enable comparisons in the presence of threats to equivalence

Although the approaches discussed so far attempt to ensure the comparability of data in cross-cultural research, they are not always successful. It is thus necessary to employ *post hoc* methods to evaluate whether the responses obtained from participants in different cultures are sufficiently comparable. Sometimes, these tests will show that the data are lacking in equivalence, in which case corrections have to be applied to the raw data. Some of the more common *post hoc* approaches will be discussed in this section.

10.3.1 Measurement invariance modeling for continuous or quasi-continuous data

Prior to conducting cross-cultural comparisons, researchers should routinely test whether the data can be meaningfully compared across cultures. We will start with the confirmatory factor analysis (CFA) approach to testing for configural, metric and scalar measurement invariance, which is applicable to continuous or quasi-continuous data and most useful when the number of groups to be compared is relatively small.

Most constructs used in cross-cultural research are sufficiently complex that multiple observed measures are needed to adequately capture the construct of interest. If multiple measures are available, a measurement analysis based on the confirmatory factor model can be conducted, provided that the data are reasonably well-behaved (i.e., there are at least five distinct response categories so that the assumption of continuity is not too grossly violated and the distribution of the data is roughly bell-shaped so that the normality assumption is somewhat

reasonable). The conventional measurement model considered here assumes that the observed responses are a reflection of the underlying construct or latent variable of interest (i.e., observed responses are a function of the posited latent variable), although unique sources of variance (measurement error) may also contribute to observed responses. For a discussion of the formative measurement model, in which the indicators are thought to cause the construct and which is not covered here, see Baumgartner and Weijters (in press) and the references cited there. Formally, the factor model can be specified as follows:

$$x^g = \tau^g + \Lambda^g \xi^g + \delta^g \qquad (1)$$

In this equation, **x** is an I x 1 vector of observed or manifest variables x_i (also called indicators), ξ is a J x 1 vector of latent variables (or common factors) ξ_j, δ is a I x 1 vector of unique factors δ_i, which are usually treated as measurement error, Λ is an I x J matrix of factor loadings Λ with typical elements λ_{ij}, which represent the strength of the relationship between the x_i and ξ_j, and τ is an I x 1 vector of equation intercepts τ_i. The superscript g refers to group g. This model differs from the usual factor model in two ways. First, it extends the single-group model to multiple groups, which is particularly useful in cross-cultural research, where researchers are often interested in studying the similarities and differences in the measurement of constructs or relationships between constructs across cultures. Second, the model includes intercepts, which are not always necessary but which are required if comparisons of means of variables and constructs are to be conducted across cultures.

The model in equation (1) is very general, but usually a more restricted version (the so-called congeneric measurement model) is considered in which each observed variable is hypothesized to load on a single factor (i.e., Λ contains only one nonzero entry per row) and the unique factors are uncorrelated (i.e., Θ is diagonal). To identify the model and set the scale of the

latent factors, one loading per factor is specified to equal one (the observed variable for which this is done is called the marker variable or reference indicator). In addition, and again for purposes of identification, the intercept of each marker variable is set to zero (in which case the means of the latent variables, which are usually called κ, can be freely estimated), or the intercepts of corresponding marker variables in different groups are set to be equal across groups and the latent factor means are restricted to equal zero in one of the groups (the reference group); the remaining latent means then express the difference in means compared to the reference group. A graphical illustration of a two-group model for two constructs, each measured by three indicators, is shown in Figure 10.1.

-- Insert Figure 10.1 about here --

The most basic requirement for conducting meaningful comparisons of constructs across cultures is that the factor structure underlying a set of observed measures be the same. This is called configural invariance. If configural invariance holds, it means that the number of common factors is the same in each culture and that a given observed variable has the same pattern of loadings on the constructs in the model (e.g., a non-zero loading on the target construct and zero loadings on non-target constructs). Configural invariance is tested by specifying the same factor model in each group and ascertaining whether this model fits adequately in each group.

If quantitative comparisons are to be conducted, stronger forms of invariance have to hold. Specifically, as explained in Steenkamp and Baumgartner (1998), if the strength of relationships between constructs is to be compared across cultures (e.g., a researcher may want to study whether the effect of attitudes on behavioral intentions is stronger in an individualistic than in a collectivistic culture, whereas the effect of subjective norms is stronger in a collectivistic than in an individualistic culture), metric invariance has to be satisfied. This means

that the factor loadings of corresponding items have to be the same across the groups to be compared (e.g., in the two-group case, $\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)}$). If, on the other hand, a researcher wants to compare the means of constructs across cultures (e.g., a researcher may want to study whether ethnocentric tendencies are stronger in a collectivistic than in an individualistic culture), then scalar invariance has to hold. This means that, in addition to the factor loadings, the intercepts of corresponding items have to be invariant across groups as well (i.e., $\lambda_{ij}^{(1)} = \lambda_{ij}^{(2)}$ and $\tau_i^{(1)} = \tau_i^{(2)}$).

It frequently happens that metric or scalar invariance hold for some of the items but not for all of them. That is, the indicators of a given construct satisfy partial but not full metric or scalar invariance. This is fine as long as metric or scalar invariance holds for at least two items per construct (see Steenkamp and Baumgartner 1998). Obviously, it is preferable if invariance holds for most, if not all, indicators of a construct, because in that case one can have greater confidence that the construct means the same, and can be measured similarly or identically, in all cultures to be compared.

To statistically test metric or scalar invariance, chi-square difference tests are usually used. For metric invariance, the fit of the model of equal loadings is compared to the fit of the configural invariance model. For scalar invariance, the fit of the model of equal loadings and intercepts is compared to the fit of the metric invariance model. If invariance of a given kind holds, the fit of the more restrictive model should not be significantly worse than the fit of the more general model. If full metric or scalar invariance are rejected, the modification indices (which indicate how much a restricted parameter hurts model fit) can be used to identify invariance constraints that have to be relaxed. In addition to chi-square difference tests, alternative fit indices are sometimes used to ascertain whether invariance of a given type holds, but unfortunately the criteria used to judge differences in fit of competing models are somewhat

arbitrary. Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000) provide

additional detail on invariance testing.

10.3.2 Measurement invariance modeling for ordinal data

Research has shown that even when the data are not strictly continuous, the conventional

confirmatory factor analysis is reasonable as long as there are at least five distinct response

categories and the spacing of the categories approximates an interval scale. If these conditions

are not met, procedures specifically designed for ordinal (including binary) scales have to be

employed. In the so-called item response theory (IRT) approach, the assumption is that the

observed response is a discretized version of an underlying continuous response, which has to be

recovered from the observed responses. If the underlying continuous response falls between a

particular lower and upper threshold, then one will observe a response in a certain scale category.

The task in IRT is to find the thresholds that mark the boundaries between the response

categories. Since the item intercepts and thresholds cannot be identified simultaneously, the

intercepts are generally set to zero. In addition to the thresholds, the model also contains slope

parameters similar to factor loadings, which are referred to as discrimination parameters in IRT

(see Baumgartner and Weijters, in press, for a discussion, as well as the references cited there).

Invariance testing is similar to the continuous case, except that one examines the equivalence of

the thresholds and slope parameters across groups (Meade & Lautenschlager, 2004; Millsap &

Yun-Tein, 2004; Reise, Widaman, & Pugh, 1993). The situation where thresholds and/or slopes

are not the same across groups is called differential item functioning in IRT.

In practice, invariance testing is more complex in the ordinal case. First, assessing model

fit is less developed in IRT modeling. For example, commonly used alternative fit indices, which

researchers generally rely upon to assess model fit heuristically, are not available in IRT. Second and most importantly, it is more difficult to assess metric and scalar invariance and to determine which items violate metric and/or scalar invariance, in part because model modification indices are unavailable. If the data can be treated as continuous, it is much simpler to use conventional invariance testing, but unfortunately this is not always possible (e.g., if the observed data are binary).

10.3.3 Measurement invariance modeling for a large number of groups and other recent extensions

In principle, it is possible to conduct invariance tests across many different groups. For example, one of the authors has tested for metric and scalar invariance of three constructs measured by a total of 25 items across 28 different countries. However, although the testing procedure can be automatized to some extent (e.g., by using the automatic model modification procedure in LISREL), it is rather cumbersome and often challenging. Furthermore, the many data-driven, sequential model modifications that will likely occur in this process raise the very real possibility that the search procedure will not identify the most appropriately constrained model and that the final model will be idiosyncratic to the data set at hand. Thus, the invariance testing procedure described in the previous section is only practicable for a relatively small number of countries or cultures (say up to 10).

As an alternative, Asparouhov and Muthén (2014) have proposed the so-called alignment method. This method consists of two steps. In the first step, a configural invariance model is estimated in which no restrictions on the loadings and intercepts are imposed but the factor means are set to zero and the factor variances to one in all groups. The restrictions on the factor

means and variances are necessary to identify the model. This model will have the same fit as the configural model described earlier. In the second step, the factor means and variances are freed, but in order to identify the model a so-called simplicity function is optimized, which minimizes the degree of non-invariance between all loadings and intercepts. Essentially, the factor means and factor variances are compared under the highest degree of measurement invariance possible. Whether or not this degree of measurement invariance is sufficient for meaningful comparisons is another question. However, when the number of groups to be compared is large and the factor model is reasonably complex, approximate measurement invariance may be the best one can hope for, and comparisons based on the alignment method should be preferable to results obtained assuming complete invariance or complete non-invariance.

Another recent extension of invariance testing is the consideration of a random effects specification for the loadings and intercepts (Asparouhov & Muthén, 2012; Muthén & Asparouhov, 2013). In the conventional invariance testing approach, the loadings and intercepts are assumed to be fixed, in the sense that a given item is assumed to have a certain loading and intercept in a particular group, which are to be estimated based on the sample at hand. If there is measurement invariance, then the loadings and intercepts will be equal in different groups. With a random effects specification, the loadings and intercepts have a certain mean and variance across all groups, the groups for which data are available are a sample of all possible groups, and the means and variances of the loadings and intercepts are estimated based on the sample of groups available. Measurement invariance implies that the variability in loadings and intercepts across all groups is small. The random effects specification leads to a two-level factor model in which there is both within-group and between-group variation in the loadings and/or intercepts. In the most general case, both the intercepts and loadings are allowed to vary across groups.

Measurement invariance no longer means that the loadings and intercepts are equal across groups, but that the measurement parameters have a common mean and variance and that the variability of the loadings and intercepts across groups is small. Again, it is not clear whether approximate measurement invariance is sufficient for meaningful comparisons across groups, but for a large number of groups a random effects specification may be an attractive model. Furthermore, as described in Muthén and Asparouhov (2013), the consideration of alternative measurement model specifications enables a host of interesting comparisons, such as random intercepts and non-random loadings that vary within and between groups, measurement invariance with equal within- and between-group loadings, and random intercepts and random loadings (De Jong & Steenkamp, 2010; De Jong, Steenkamp, & Fox, 2007; Muthén & Asparouhov, 2013).

In summary, several important developments have occurred in the literature on invariance testing in recent years. Although practical applications of these techniques are still uncommon, the increased access to large data sets from many different cultures and the ready availability of computer programs to estimate these models makes it likely that these approaches will be used more in cross-cultural research in the near future.

10.3.4 Individual-level correction procedures

The focus of invariance testing is on ascertaining whether the data are suitable for cross-cultural comparisons. Subsequent analyses, in which construct means or relationships between constructs are compared across cultures, can then be based on models that are appropriately constrained across groups (e.g., only items that actually exhibit metric or scalar invariance are restricted to have invariant loadings or intercepts, so model misspecifications can be avoided).

Although invariance testing does not correct for violations of measurement equivalence *per se*, cross-cultural comparisons at the construct level are based on the items that are invariant across cultures, so the comparisons are valid. However, one important prerequisite is that a sufficient number of items be cross-culturally comparable, as explained earlier, otherwise comparisons of interest may not be justified.

An alternative is to assess potential causes of non-equivalence and correct for them explicitly. Two types of biasing effects have been discussed in the literature (Baumgartner & Weijters, 2015). On the one hand, there are systematic response tendencies that are more or less independent of the content of the substantive items that the researcher is interested in. The most important of these response styles are various systematic scale usage differences such as (dis)acquiescent, extreme and midpoint responding (Baumgartner & Steenkamp, 2001). On the other hand, there are systematic response tendencies that depend on the content of the substantive items but do not accurately reflect what the researcher is trying to measure with these items. The most well-known of these responses biases is socially desirable responding (SDR), where people's responses are motivated by a desire to present a favorable image of oneself (Steenkamp et al., 2010).

Specialized scales are needed to assess SDR, but scale usage differences due to (dis)acquiescent, extreme and midpoint responding are often measured based on the substantive items themselves. This is attractive because no additional questions have to be included in the survey. A popular correction procedure is within-person standardization (or mean-centering) of the data across (preferably) many different items, for example, all the items to be compared across cultures (Fischer, 2004). More sophisticated versions of this general idea have been proposed as well (Rossi, Gilula, & Allenby, 2001). In general, we advise against the use of these

methods because substance and style cannot be clearly separated when the substantive items themselves are used to assess differences in scale usage. However, there is one important exception to this recommendation. If content-based responding is different from stylistic responding, then basing response style measures on the substantive items may be justified. For example, if there are both regular and reversed items, ARS can be assessed based on the number of agreement responses to both regular and reversed items before recoding the reversed items. For instance, in a scale measuring consumer ethnocentrism, a respondent high in ethnocentrism should disagree with an item such as "I like buying foreign products" if the response were driven by substantive considerations, but might agree with the same item if the response were driven by acquiescence. Thus, substantive and stylistic responding can be distinguished even though the same item is used to assess both. Unfortunately, even this method may be problematic in cross-cultural research because, as discussed earlier, research shows that some cultures (e.g., collectivistic cultures) may respond to reversed items differently than other cultures (e.g., individualistic cultures). In general, independent control items should be used to assess and correct for scale usage differences on substantive items (see the next section for a more extended discussion).

Regardless of whether one wants to control for systematic scale usage differences or SDR, the approach is the same. First, the systematic response tendency that is hypothesized to bias people's substantive responses has to be measured (based on independent control items or an SDR scale). Second, the observed responses are purified by regressing them on the measure of systematic response tendencies and the residuals from this regression are then used in subsequent analyses. Alternatively, the measure of systematic response tendencies can be included as a control variable in the analysis of interest. As discussed in Podsakoff et al. (2003),

it is best to do the correction at the individual item level. One disadvantage of using an overall measure of systematic response tendencies is that measurement error in the assessment of systematic response tendencies is not taken into account. If this is important, then the Representative Indicators Response Style Means and Covariance Structure (RIRSMACS) approach suggested by Weijters, Schillewaert, and Geuens (2008) can be used. In this case, multiple indicators of a given response style (or other systematic response tendencies of interest) are included in the model and a confirmatory factor analysis is used in which each individual indicator is related to both the underlying substantive factor and the response style factor, which is measured by multiple items. It is also possible to include multiple biasing influences if appropriate measures are available. Recently, several authors have also suggested more complex approaches that do not assume linear effects of the response styles (Kankaraš, Moors, & Vermunt, 2010; Morren, Gelissen, & Vermunt, 2011).

The correction procedures described in this section assess stylistic response tendencies at the individual-respondent level, and the control for systematic response biases also occurs at the respondent level. If there are systematic differences in biased responding across cultures, then cross-cultural differences are taken into account implicitly as well, although the correction is at the individual-respondent level. Since it is likely that there are both individual-level and culture-level determinants of systematic response tendencies, it is advantageous to control for both. However, if response biases are thought to occur primarily between different cultures, it may be sufficient to control for systematic response tendencies at the group-level only. Such an approach is described next.

10.3.5 Group-level correction procedures

Weijters, Baumgartner, et al. (2016) recently proposed a technique called the calibrated sigma method, which corrects for scale usage differences at the group level. The approach is an extension of the sigma method suggested by Likert (1932), in which he introduced what is now known as Likert scaling. The idea is simple. Assume that a researcher uses a 5-point scale with response categories of strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree to assess the extent of people's (dis)agreement with a series of statements designed to measure a construct of interest. Normally, consecutive integers ranging from 1 to 5 (or maybe -2 to +2) are used to code people's responses. This coding assumes that respondents treat the five response categories as an interval scale and, more importantly and probably more questionably, that respondents from different countries understand and use the scale in the same way. Likert noted that the numbers assigned to people's responses need not be equally spaced, and that different sets of numbers could be assigned to scale responses. Weijters et al. further proposed that the weights used to weight people's responses should be based on independent control items which are (a) unrelated to the substantive constructs being measured and (b) unrelated to each other. The first requirement ensures that scale usage differences are not confounded with substantive differences; the second requirement ensures that the weights derived from the control items reflect pure scale usage differences, not substantive differences based on what the control items have in common.

The calibrated sigma method involves the following steps. First, the questionnaire has to contain a number of control items that are heterogeneous in meaning. For example, the 16-item scale proposed by Greenleaf (1992) to measure extreme responding may be used for this purpose (although it should be noted that little evidence is available that this scale is equally appropriate across cultures). Cross-cultural researchers may be hesitant to include so many control items in

their surveys, when the purpose of these items is solely to control for differences in scale usage, but unfortunately this is necessary unless prior research has shown that the substantive items used in the survey are free of scale usage differences in the cultures of interest (which is rarely the case). Second, the proportion of responses in response category k is computed across all control items and all respondents within group g. Third, the proportions are converted into normal scores. Fourth, the original responses are weighted in a group-specific way using the normal scores obtained in the previous step (see Weijters et al. for a worked example of the method and additional detail). Weijters et al. report an illustration in which they compared Dutch- and French-speaking respondents on the construct of Need for Predictability. When participants' responses were coded 1 to 5, the French-speaking sample had a higher mean Need for Predictability than the Dutch-speaking sample. However, when the raw scores were re-weighted using the calibrated sigma method, the difference vanished, consistent with theoretical expectations that there should be no difference in Need for Predictability between Dutch- and French-speaking Belgians.

The calibrated sigma method has the following advantages. First, since the correction is done at the group level and the normal scores used to re-weight the raw scores are computed across many items and respondents, the procedure should be highly reliable. In contrast, if the correction is done at the individual level, it may not be very accurate. Second, the method is very easy to use. Third, it can be applied even when the number of cultures to be compared is large (say, more than 20). Fourth, it is effective even when systematic bias is uniform across items, in which case measurement invariance testing would not be able to detect the bias. Fifth, the weights may be based on previous studies or a subset of respondents, so that the control items

need not be administered to the entire sample, which is the most serious drawback of the procedure.

10.4 Conclusion

In this chapter we discussed various methodological issues that commonly arise in cross-cultural research. The overriding concern is usually how to conduct meaningful comparisons across cultures. We discussed both *a priori* questionnaire design principles based on the psychology of survey response that researchers should consider before they collect data and *post hoc* data analysis strategies that should be used when responses are actually compared across cultures. Both approaches should be routinely applied when research involves multiple cultures, and we hope that following these guidelines will contribute to greater validity of cross-cultural comparisons.

References

Asparouhov, T., & Muthén, B. (2012). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495-508.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*, 491-509.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*(May), 143-156.

Baumgartner, H., & Weijters, B. (2015). Response Biases in Cross-Cultural Measurement. *Handbook of Culture and Consumer Behavior*, 150.

Baumgartner, H., & Weijters, B. (in press). Measurement Models for Marketing Constructs. In B. Wierenga & R. van der Lans (Eds.), *Handbook of Marketing Decision Models (New Edition)*: Springer.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185-216.

Brislin, R. W. (1986). Research instruments. *Field methods in cross-cultural research*, 159-162.

Cabooter, E., Millet, K., Weijters, B., & Pandelaere, M. (2016). The 'I'in Extreme Responding. *Journal of Consumer Psychology, forthcoming*.

Chen, C., Lee, S.-y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 170-175.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Sociology, 40*.

Dawar, N., & Parker, P. (1994). Marketing universals: Consumers' use of brand name, price, physical appearance, and retailer reputation as signals of product quality. *The Journal of Marketing*, 81-95.

De Jong, M. G., Fox, J.-P., & Steenkamp, J.-B. E. M. (2015). Quantifying Under-and Overreporting in Surveys Through a Dual-Questioning-Technique Design. *Journal of Marketing Research, 52*(6), 737-753.

De Jong, M. G., Pieters, R., & Fox, J.-P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research, 47*(1), 14-27.

De Jong, M. G., & Steenkamp, J.-B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika, 75*(1), 3-32.

De Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing Measurement Invariance in Cross-national Consumer Research Using a Hierarchical IRT Model. *Journal of Consumer Research, 34*(22), 260-278.

De Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research, 45*(February), 104-115.

de Langhe, B., Puntoni, S., Fernandes, D., & van Osselaer, S. M. J. (2011). The Anchor Contraction Effect in International Marketing Research. *Journal of Marketing Research, 48*(2), 366-380.

Douglas, S. P., & Craig, C. S. (2007). Collaborative and iterative translation: An alternative approach to back translation. *Journal of International Marketing, 15*(1), 30-43.

Fischer, R. (2004). Standardization to account for cross-cultural response bias - A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology, 35*(3), 263-282. doi: 10.1177/0022022104264122

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*(2), 303-315.

Gardner, W. L., Gabriel, S., & Lee, A. Y. (1999). "I" value freedom, but "we" value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological Science, 10*(4), 321-326.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (Vol. 3, pp. 225-242). New York: Seminar Press,.

Haberstroh, S., Oyserman, D., Schwarz, N., Kühnen, U., & Ji, L.-J. (2002). Is the interdependent self more sensitive to question context than the independent self? Self-construal and the observation of conversational norms. *Journal of Experimental Social Psychology, 38*(3), 323-329.

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual differences, 44*(4), 932-942.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., . . . Smith, T. W. (2010). Comparative survey methodology. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, 1-16.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010). Designing questionnaires for multipopulation research. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, 31-57.

Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In Presser, S. et al. (Eds), *Methods for testing and evaluating survey questionnaires, New York: Wiley,* 453-473.

Harkness, J. A., & Schoua-Glusberg, A. S. (1998). Questionnaires in translation. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence* (pp.87-126). ZUMA-Nachrichten Spezial, 3, Mannheim, Germany: ZUMA

Harzing, A.-W. (2006). Response Styles in Cross-national Survey Research A 26-country Study. *International Journal of Cross Cultural Management, 6*(2), 243-266.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296-309.

Hupka, R. B., Zaleski, Z., Otto, J., Reidl, L., & Tarabrina, N. V. (1997). The colors of anger, envy, fear, and jealousy a cross-cultural study. *Journal of Cross-Cultural Psychology, 28*(2), 156-171.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The Relation between Culture and Response Styles: Evidence from 19 Countries. *Journal of Cross-Cultural Psychology, 36*(2), 264-277.

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for measurement invariance with latent class analysis. *Cross-cultural analysis: Methods and applications*, 359-384.

Laczniak, R. N. (2015). The Journal of Advertising and the Development of Advertising Theory: Reflections and Directions for Future Research. *Journal of Advertising, 44*(4), 429-433.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology, 90*(1), 165.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology.*

Luchs, M. G., Naylor, R. W., Irwin, J. R., & Raghunathan, R. (2010). The Sustainability Liability: Potential Negative Effects of Ethicality on Product Preference. *Journal of Marketing, 74*(5), 18-31.

Marín, G., Gamba, R. J., & Marìn, B. V. (1992). Extreme response styles and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology, 23*(December), 498-509.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review, 98*(2), 224.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361-388.

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

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2011). Dealing with Extreme Response Style in Cross-Cultural Resarch: A Restricted Latent Class Factor Analysis Approach. *Sociological Methodology, 41*(1), 13-47.

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2012). Exploring the Response Process of Culturally Differing Survey Respondents with a Response Style: A Sequential Mixed Methods Study. *Field Methods, 25*(2), 162-181.

Muthén, B., & Asparouhov, T. (2013). New methods for the study of measurement invariance with many groups. *Mplus. statmodel. com [12.04. 2014]*.

Okazaki, S., & Mueller, B. (2007). Cross-cultural advertising research: where we have been and where we need to go. *International Marketing Review, 24*(5), 499-518.

Oyserman, D., & Lee, S. W. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin, 134*(2), 311.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes*: Academic Press.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: a critical review and meta-analysis. *Psychological Bulletin, 114*(2), 363.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach. *Journal of the American Statistical Association, 96*(March), 20-31.

Salancik, G. R. (1984). On priming, consistency, and order effects in job attitude assessment: With a note on current research.

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

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. *Survey methods in multinational, multiregional and multicultural contexts*, 177-190.

Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly, 55*(1), 3-23.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*(1), 50-61.

Smith, T. W. (2004). Developing and evaluating cross-national survey instruments. In R. M. Groves, G. Kalton, J. N. K. Rao, N. Schwarz & C. Skinner (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 431-452).

Smith, T. W., Mohler, P. P., Harkness, J., & Onodera, N. (2005). Methods for assessing and calibrating response scales across countries and languages. *Comparative sociology, 4*(3), 365.

Steenkamp, J.-B. E. M. (2001). The role of national culture in international marketing research. *International Marketing Review, 18*(1), 30-44.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*(June), 78-90.

Steenkamp, J.-B. E. M., De Jong, M. G., & Baumgartner, H. (2010). Socially Desirable Response Tendencies in Survey Research. *Journal of Marketing Research, 47*(2), 199-214.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin, 103*(3), 299-314.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*: Cambridge University Press.

Uskul, A. K., Oyserman, D., & Schwarz, N. (2010). Cultural emphasis on honor, modesty or self-enhancement: Implications for the survey response process. In J. Harkness et al. (Eds.), *Survey methods in multinational, multiregional and multicultural contexts* (pp. 191-201). New York: Wiley.

Uskul, A. K., Oyserman, D., Schwarz, N., Lee, S. W., & Xu, A. J. (2013). How successful you have been in life depends on the response scale used: The role of cultural mindsets in pragmatic inferences drawn from question format. *Social Cognition, 31*(2), 222.

Usunier, J.-C., can Herk, H, & Lee, J. A. (2017). International and Cross-Cultural Business Research: Sage.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, 1(2), 89-99.

Van de Vijver, F. J., & Leung, K. (1997). *Methods and Data Analysis for Cross-cultural Research* (Vol. 1): Sage.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology, 35*(3), 346-360.

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research, 47*(1), 157-172.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.

Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research, 49*(5), 737-747.

Weijters, B., Baumgartner, H., & Geuens, M. (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, 944-960.*

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed Item Bias: An Integrative Model. *Psychological Methods, 18*(3), 320-334.

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The Effect of Familiarity with the Response Category Labels on Item Response to Likert Scales. *Journal of Consumer Research, 40*(2), 368-381.

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

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409–422.

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. *Journal of Consumer Research, 30*(1), 72-91. doi: 10.1086/374697

Figure 10.1