Capturing bias in structural equation modeling

van de Vijver, F.J.R.

Published in:
Cross-cultural analysis. Methods and applications

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

1

Capturing Bias in Structural Equation Modeling

Fons J. R. van de Vijver
Tilburg University and North-West University

1.1 INTRODUCTION

Equivalence studies are coming of age. Thirty years ago there were few conceptual models and statistical techniques to address sources of systematic measurement error in cross-cultural studies (for early examples, see Clearly & Hilton, 1968; Lord, 1977, 1980; Poortinga, 1971). This picture has changed; in the last decades conceptual models and statistical techniques have been developed and refined. Many empirical examples have been published. There is a growing awareness of the importance in the field for the advancement of cross-cultural theorizing. An increasing number of journals require authors who submit manuscripts of cross-cultural studies to present evidence supporting the equivalence of the study measures. Yet, the burgeoning of the field has not led to a convergence in conceptualizations, methods, and analyses. For example, educational testing focuses on the analysis of items as sources of problems of cross-cultural comparisons, often using item response theory (e.g., Emenogu & Childs, 2005). In personality psychology, exploratory factor analysis is commonly applied as a tool to examine the similarity of factors underlying a questionnaire (e.g., McCrae, 2002). In survey research and marketing, structural equation modeling (SEM) is most frequently employed (e.g., Steenkamp & Baumgartner, 1998). From a theoretical perspective, these models are related; for example, the relationship of item response theory and confirmatory factor analysis (as derived from a general latent variable model) has been described by Brown (2006). However, from a practical

perspective, the models can be seen as relatively independent paradigms; there are no recent studies in which various bias models are compared (an example of an older study in which procedures are compared that are no longer used has been described by Shepard, Camilli, & Averill, 1981).

In addition to the diversity in mathematical developments, conceptual frameworks for dealing with cross-cultural studies have been developed in cross-cultural psychology, which, again, have a slightly different focus. It is fair to say that the field of equivalence is still expanding in both conceptual and statistical directions and that rapprochement of the approaches and best practices that are broadly accepted across various fields are not just around the corner.

The present chapter relates the conceptual framework about measurement problems that is developed in cross-cultural psychology (with input from various other sciences studying cultures and cultural differences) to statistical developments and current practices in SEM vis-à-vis multigroup testing. More specifically, I address the question of the strengths and weaknesses of SEM from a conceptual bias and equivalence framework. There are few publications in which more conceptually based approaches to bias that are mainly derived from substantive studies are linked to more statistically based approaches such as developed in SEM. This chapter adds to the literature by linking two research traditions that have worked largely independently in the past, despite the overlap in bias issues addressed in both traditions. The chapter deals with the question to what extent the study of equivalence, as implemented in SEM, can address all the relevant measurement issues of cross-cultural studies. The first part of the chapter describes a theoretical framework of bias and equivalence. The second part describes various procedures and examples to identify bias and address equivalence. The third part discusses the identification of all the bias types distinguished using SEM. The fourth part presents a SWOT analysis (strengths, weaknesses, opportunities, and threats) of SEM in dealing with bias sources in cross-cultural studies. Conclusions are drawn in the final part.

1.2 BIAS AND EQUIVALENCE

The bias framework is developed from the perspective of cross-cultural psychology and attempts to provide a comprehensive taxonomy of all

systematic sources of error that can challenge the inferences drawn from cross-cultural studies (Poortinga, 1989; Van de Vijver & Leung, 1997). The equivalence framework addresses the statistical implications of the bias framework and defines conditions that have to be fulfilled before inferences can be drawn about comparative conclusions dealing with constructs or scores in cross-cultural studies.

1.2.1 Bias

Bias refers to the presence of nuisance factors (Poortinga, 1989). If scores are biased, the meaning of test scores varies across groups and constructs and/or scores are not directly comparable across cultures. Different types of bias can be distinguished (Van de Vijver & Leung, 1997).

1.2.1.1 Construct Bias

There is *construct bias* if a construct differs across cultures, usually due to an incomplete overlap of construct-relevant behaviors. An empirical example can be found in Ho's (1996) work on filial piety (defined as a psychological characteristic associated with being "a good son or daughter"). The Chinese concept, which includes the expectation that children should assume the role of caretaker of elderly parents, is broader than the Western concept.

1.2.1.2 Method Bias

Method bias is the generic term for all sources of bias due to factors often described in the methods section of empirical papers. Three types of method bias have been defined, depending on whether the bias comes from the sample, administration, or instrument. Sample bias refers to systematic differences in background characteristics of samples with a bearing on the constructs measured. Examples are differences in educational background that can influence a host of psychological variables such as cognitive tests. Administration bias refers to the presence of cross-cultural conditions in testing conditions, such as ambient noise. The potential influence of interviewers and test administrators can also be mentioned here. In cognitive testing, the presence of the tester does not need to be obtrusive (Jensen, 1980). In survey research there is more evidence for interviewer effects (Lyberg et al., 1997). Deference to the interviewer has been reported; participants are more likely to display positive attitudes to

an interviewer (e.g., Aquilino, 1994). Instrument bias is a final source of bias in cognitive tests that includes instrument properties with a pervasive and unintended influence on cross-cultural differences such as the use of response alternatives in Likert scales that are not identical across groups (e.g., due to a bad translation of item anchors).

1.2.1.3 Item Bias

Item bias or differential item functioning refers to anomalies at the item level (Camilli & Shepard, 1994; Holland & Wainer, 1993). According to a definition that is widely used in education and psychology, an item is biased if respondents from different cultures with the same standing on the underlying construct (e.g., they are equally intelligent) do not have the same mean score on the item. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993; Sireci, in press; Van de Vijver & Leung, 1997, in press).

Item bias can arise in various ways, such as poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, and the influence of culture-specific nuisance factors or connotations associated with the item wording. Suppose that a geography test is administered to pupils in all EU countries that ask for the name of the capital of Belgium. Belgian pupils can be expected to show higher scores on the item than pupils from other EU countries. The item is biased because it favors one cultural group across all test score levels.

1.2.2 Equivalence

Bias has implications for the comparability of scores (e.g., Poortinga, 1989). Depending on the nature of the bias, four hierarchically nested types of equivalence can be defined: construct, structural or functional, metric (or measurement unit), and scalar (or full score) equivalence. These four are further described below.

1.2.2.1 Construct Inequivalence

Constructs that are inequivalent lack a shared meaning, which precludes any cross-cultural comparison. In the literature, claims of construct

inequivalence can be grouped into three broad types, which differ in the degree of inequivalence (partial or total). The first and strongest claim of inequivalence is found in studies that adopt a strong emic, relativistic viewpoint, according to which psychological constructs are completely and inseparably linked to their natural context. Any cross-cultural comparison is then erroneous as psychological constructs are cross-culturally inequivalent.

The second type is exemplified by psychological constructs that are associated with specific cultural groups. The best examples are culture-bound syndromes. A good example is Amok, which is specific to Asian countries like Indonesia and Malaysia. Amok is characterized by a brief period of violent aggressive behavior among men. The period is often preceded by an insult and the patient shows persecutory ideas and automatic behaviors. After this period, the patient is usually exhausted and has no recollection of the event (Azhar & Varma, 2000). Violent aggressive behavior among men is universal, but the combination of triggering events, symptoms, and lack of recollection is culture-specific. Such a combination of universal and culture-specific aspects is characteristic for culture-bound syndromes. Taijin Kyofusho is a Japanese example (Suzuki, Takei, Kawai, Minabe, & Mori, 2003; Tanaka-Matsumi & Draguns, 1997). This syndrome is characterized by an intense fear that one's body is discomforting or insulting for others by its appearance, smell, or movements. The description of the symptoms suggests a strong form of a social phobia (a universal), which finds culturally unique expressions in a country in which conformity is a widely shared norm. Suzuki et al. (2003) argue that most symptoms of Taijin Kyofusho can be readily classified as social phobia, which (again) illustrates that culture-bound syndromes involve both universal and culture-specific aspects.

The third type of inequivalence is empirically based and found in comparative studies in which the data do not show any evidence for construct comparability; inequivalence here is a consequence of the lack of cross-cultural comparability. Van Leest (1997) administered a standard personality questionnaire to mainstream Dutch and Dutch immigrants. The instrument showed various problems, such as the frequent use of colloquialisms. The structure found in the Dutch mainstream group could not be replicated in the immigrant group.

1.2.2.2 Structural or Functional Equivalence

An instrument administered in different cultural groups shows structural equivalence if it measures the same construct(s) in all these groups (it should be noted that this definition is different from the common definition of structural equivalence in SEM; in a later section I return to this confusing difference in definitions). Structural equivalence has been examined for various cognitive tests (Jensen, 1980), Eysenck's Personality Questionnaire (Barrett, Petrides, Eysenck, & Eysenck, 1998), and the five-factor model of personality (McCrae, 2002). Functional equivalence as a specific type of structural equivalence refers to identity of nomological networks (Cronbach & Meehl, 1955). A questionnaire that measures, say, openness to new cultures shows functional equivalence if it measures the same psychological constructs in each culture, as manifested in a similar pattern of convergent and divergent validity (i.e., nonzero correlations with presumably related measures and zero correlations with presumably unrelated measures). Tests of structural equivalence are applied more often than tests of functional equivalence. The reason is not statistical. With advances in statistical modeling (notably path analysis as part of SEM), tests of the cross-cultural similarity of nomological networks are straightforward. However, nomological networks are often based on a combination of psychological scales and background variables, such as socioeconomic status, education, and sex. The use of psychological scales to validate other psychological scales can lead to an infinite regression in which each scale in the network that is used to validate the target construct requires validation itself. If this issue has been dealt with, the statistical testing of nomological networks can be done in path analyses or MIMIC model (*M*ultiple *I*ndicators, *Mult*iple *C*auses; Jöreskog & Goldberger, 1975), in which the background variables predict a latent factor that is measured by the target instrument as well as the other instruments studied to address the validity of the target instrument.

1.2.2.3 Metric or Measurement Unit Equivalence

Instruments show metric (or measurement unit) equivalence if their measurement scales have the same units of measurement, but a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the

same measurement units in each culture). Metric equivalence is found when a source of bias creates an offset in the scale in one or more groups, but does not affect the relative scores of individuals within each cultural group. For example, social desirability and stimulus familiarity influence questionnaire scores more in some cultures than in others, but they may influence individuals within a given cultural group in a fairly homogeneous way.

1.2.2.4 Scalar or Full Score Equivalence

Scalar equivalence assumes an identical interval or ratio scale in all cultural groups. If (and only if) this condition is met, direct cross-cultural comparisons can be made. It is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different or equal.

1.3 BIAS AND EQUIVALENCE: ASSESSMENT AND APPLICATIONS

1.3.1 Identification Procedures

Most procedures to address bias and equivalence only require cross-cultural data with a target instrument as input; there are also procedures that rely on data obtained with additional instruments. The procedures using additional data are more open, inductive, and exploratory in nature, whereas procedures that are based only on data with the target instrument are more closed, deductive, and hypothesis testing. An answer to the question of whether additional data are needed, such as new tests or other ways of data collection such as cognitive pretesting, depends on many factors. Collecting additional data is the more laborious and time-consuming way of establishing equivalence that is more likely to be used if fewer cross-cultural data with the target instrument are available; the cultural and linguistic distance between the cultures in the study are larger, fewer theories about the target construct are available, or when the need is more felt to develop a culturally appropriate measure (possibly with culturally specific parts).

1.3.1.1 Detection of Construct Bias and Construct Equivalence

The detection of construct bias and construct equivalence usually requires an exploratory approach in which local surveys, focus group discussions, or in-depth interviews are held with members of a community are used to establish which attitudes and behaviors are associated with a specific construct. The assessment of method bias also requires the collection of additional data, alongside the target instrument. Yet, a more guided search is needed than in the assessment of construct bias. For example, examining the presence of sample bias requires the collection of data about the composition and background of the sample, such as educational level, age, and sex. Similarly, identifying potential influence of cross-cultural differences in response styles requires their assessment. If a bipolar instrument is used, acquiescence can be assessed by studying the levels of agreement with both the positive and negative items; however, if a unipolar instrument is used, information about acquiescence should be derived from other measures. Item bias analyses are based on closed procedures; for example, scores on items are summed and the total score is used to identify groups in different cultures with a similar performance. Item scores are then compared in groups with a similar performance from different cultures.

1.3.1.2 Detection of Structural Equivalence

The assessment of structural equivalence employs closed procedures. Correlations, covariances, or distance measures between items or subtests are used to assess their dimensionality. Coordinates on these dimensions (e.g., factor loadings) are compared across cultures. Similarity of coordinates is used as evidence in favor of structural equivalence. The absence of structural equivalence is interpreted as evidence in favor of construct inequivalence. Structural equivalence techniques, as they are closed procedures, are helpful to determine the cross-cultural similarity of constructs, but they may need to be complemented by open procedures, such as focus group discussions to provide a comprehensive coverage of the definition of construct in a cultural group. Functional equivalence, on the other hand, is based on a study of the convergent and divergent validity of an instrument measuring a target construct. Its assessment is based on open procedures, as additional instruments are required to establish this validity.

1.3.1.3 Detection of Metric and Scalar Equivalence

Metric and scalar equivalence are also on closed procedures. SEM is often used to assess relations between items or subtests and their underlying constructs. It can be concluded that open and closed procedures are complementary.

1.3.2 Examples

1.3.2.1 Examples of Construct Bias

An interesting study of construct bias has been reported by Patel, Abas, Broadhead, Todd, and Reeler (2001). These authors were interested how depression is expressed in Zimbabwe. In interviews with Shona speakers, they found that

> Multiple somatic complaints such as headaches and fatigue are the most common presentations of depression. On inquiry, however, most patients freely admit to cognitive and emotional symptoms. Many somatic symptoms, especially those related to the heart and the head, are cultural metaphors for fear or grief. Most depressed individuals attribute their symptoms to "thinking too much" (kufungisisa), to a supernatural cause, and to social stressors. Our data confirm the view that although depression in developing countries often presents with somatic symptoms, most patients do not attribute their symptoms to a somatic illness and cannot be said to have "pure" somatisation. (p. 482)

This conceptualization of depression is only partly overlapping with western theories and models. As a consequence, western instruments will have a limited suitability, particularly with regard to the etiology of the syndrome.

There are few studies that are aimed at demonstrating construct inequivalence, but studies have found that the underlying constructs were not (entirely) comparable and hence, found evidence for construct inequivalence. For example, De Jong and colleagues (2005) examined the cross-cultural construct equivalence of the Structured Interview for Disorders to of Extreme Stress (SIDES), an instrument designed to assess symptoms of Disorders of Extreme Stress Not Otherwise Specified (DESNOS). The interview aims to measure the psychiatric sequelae of interpersonal victimization, notably the consequences of war, genocide, persecution, torture,

and terrorism. The interview covers six clusters, each with a few items; examples are alterations in affect regulation and impulses. Participants completed the SIDES as a part of an epidemiological survey conducted between 1997 and 1999 among large samples of survivors of war or mass violence in Algeria, Ethiopia, and Gaza. Exploratory factor analyses were conducted for each of the six clusters; the cross-cultural equivalence of the six clusters was tested in a multisample, confirmatory factor analysis. The Ethiopian sample was sufficiently large to be split up into two subsamples. Equivalence across these subsamples was supported. However, comparisons of this model across countries showed a very poor fit. The authors attributed this lack of equivalence to the poor applicability of various items in these cultural contexts; they provide an interesting table in which they compare the prevalence of various symptoms in these populations with those in field trials to assess Post-Traumatic Stress Disorder that are included in the *DSM–IV* of the American Psychiatric Association. The general pattern was that most symptoms were less prevalent in these three areas than reported in the manual and that there were also large differences in prevalence across the three areas. Findings indicated that the factor structure of the SIDES was not stable across samples; thus construct equivalence was not shown. It is not surprising that items with such large cross-cultural differences in endorsement rates are not related in a similar manner across cultures. The authors conclude that more sensitivity for the cultural context and the cultural appropriateness of the instrument would be needed to compile instruments that would be better able to stand cross-cultural validation. It is an interesting feature of the study that the authors illustrate how this could be done by proposing a multistep interdisciplinary method that accommodates universal chronic sequelae of extreme stress and accommodates culture-specific symptoms across a variety of cultures. The procedure illustrates how constructs with only a partial overlap across cultures require a more refined approach to cross-cultural comparisons as shared and unique aspects have to be separated. It may be noted that this approach exemplifies universalism in cross-cultural psychology (Berry et al., 2002), according to which the core of psychological constructs tends to be invariant across cultures but manifestations may take culture-specific forms.

As another example, it has been argued that organizational commitment contains both shared and culture-specific components. Most western research is based on a three-componential model (e.g., Meyer &

Allen, 1991; cf. Van de Vijver & Fischer, 2009) that differentiates between affective, continuance, and normative commitment. Affective commitment is the emotional attachment to organizations, the desire to belong to the organization and identification with the organizational norms, values, and goals. Normative commitment refers to a feeling of obligation to remain with the organization, involving normative pressure and perceived obligations by significant others. Continuance commitment refers to the costs associated with leaving the organization and the perceived need to stay. Wasti (2002) argued that continuance commitment in more collectivistic contexts such as Turkey, loyalty and trust are important and strongly associated with paternalistic management practices. Employers are more likely to give jobs to family members and friends. Employees hired in this way will show more continuance commitment. However, Western measures do not address this aspect of continuance commitment. A meta-analysis by Fischer and Mansell (2007) found that the three components are largely independent in Western countries, but are less differentiated in lower income contexts. These findings suggest that the three components become more independent with increasing economic affluence.

1.3.2.2 Examples of Method Bias

Method bias has been addressed in several studies. Fernández and Marcopulos (2008) describe how incomparability of norm samples made international comparisons of the Trail Making Test (an instrument to assess attention and cognitive flexibility) impossible: "In some cases, these differences are so dramatic that normal subjects could be classified as pathological and vice versa, depending upon the norms used" (p. 243). Sample bias (as a source of method bias) can be an important rival hypothesis to explain cross-cultural score differences in acculturation studies. Many studies compare host and immigrant samples on psychological characteristics. However, immigrant samples that are studied in Western countries often have lower levels of education and income than the host samples. As a consequence, comparisons of raw scores on psychological instruments may be confounded by sample differences. Arends-Tóth and Van de Vijver (2008) examined similarities and differences in family support in five cultural groups in the Netherlands (Dutch mainstreamers, Turkish-, Moroccan-, Surinamese-, and Antillean-Dutch). In each group, provided

support was larger than received support, parents provided and received more support than siblings, and emotional support was stronger than functional support. The cultural differences in mean scores were small for family exchange and quality of relationship, and moderate for frequency of contact. A correction for individual background characteristics (notably age and education) reduced the effect size of cross-cultural differences from .04 (proportion of variance accounted for by culture before correction) to .03 (after correction) for support and from .07 to .03 for contact. So, it was concluded that the cross-cultural differences in raw scores were partly unrelated to cultural background and had to be accounted for by background characteristics.

The study of response styles (and social desirability that is usually not viewed as a style, but also involves self-presentation tactics) enjoys renewed interest in cross-cultural psychology. In a comparison of European countries, Van Herk, Poortinga, and Verhallen (2004) found that Mediterranean countries, particularly Greece, showed higher acquiescent and extreme responding than Northwestern countries in surveys on consumer research. They interpreted these differences in terms of the individualism versus collectivism dimension. In a meta-analysis across 41 countries, Fischer, Fontaine, Van de Vijver, and Van Hemert (in press) calculated acquiescence scores for various scales in the personality, social psychological, and organizational domains. A small but significant percentage (3.1%) of the overall variance was shared among all scales, pointing to a systematic influence of response styles in cross-cultural comparisons. In presumably the largest study of response styles, Harzing (2006) found consistent cross-cultural differences in acquiescence and extremity responding across 26 countries. Cross-cultural differences in response styles are systematically related to various country characteristics. Acquiescence and extreme responding are more prevalent in countries with higher scores on Hofstede's collectivism and power distance, and GLOBE's uncertainty avoidance. Furthermore, extraversion (at the country level) is a positive predictor of acquiescence and extremity scoring. Finally, she found that English-language questionnaires tend to evoke less extremity scoring and that answering items in one's native language is associated with more extremity scoring. Cross-cultural findings on social desirability also point to the presence of systematic differences in that more affluent countries show, on average, lower scores on social desirability (Van Hemert, Van de Vijver, Poortinga, & Georgas, 2002).

Instrument bias is a common source of bias in cognitive tests. An example can be found in Piswanger's (1975) application of the Viennese Matrices Test (Formann & Piswanger 1979). A Raven-like figural inductive reasoning test was administered to high school students in Austria, Nigeria, and Togo (educated in Arabic). The most striking findings were the cross-cultural differences in item difficulties related to identifying and applying rules in a horizontal direction (i.e., left to right). This was interpreted as bias in terms of the different directions in writing Latin-based languages as opposed to Arabic.

1.3.2.3 Examples of Item Bias

More studies of item bias have been published than of any other form of bias. All widely used statistical techniques have been used to identify item bias. Item bias is often viewed as an undesirable item characteristic that should be eliminated. As a consequence, items that are presumably biased are eliminated prior to the cross-cultural comparisons of scores. However, it is also possible to view item bias as a source of cross-cultural differences that is not to be eliminated but requires further examination (Poortinga & Van der Flier, 1988). The background of this view is that item bias, which by definition involves systematic cross-cultural differences, can be interpreted as referring to culture-specifics. Biased items provide information about cross-cultural differences on other constructs than the target construct. For example in a study on intended self-presentation strategies by students in job interviews involving 10 countries, it was found that the dress code yielded biased items (Sandal et al., in preparation). Dress code was an important aspect of self-presentation in more traditional countries (such as Iran and Ghana) whereas informal dress was more common in more modern countries (such as Germany and Norway). These items provide important information about self-presentation in these countries, which cannot be dismissed as bias but that should be eliminated.

The more than 40 years of item bias research after Cleary and Hilton's (1968) first study have not led to accumulated insights as to which items tend to be biased. In fact, one of the complaints has been the lack of accumulation. Educational testing has been an important domain of application of item bias. Linn (1993), in a review of the findings, came to the sobering conclusion that no general findings have emerged about which item characteristics are associated with item bias; he argued that item

difficulty was the only characteristic that was more or less associated with bias. The item bias tradition has not led to widely accepted practices about item writing for multicultural assessment. One of the problems in accumulating knowledge from the item bias tradition about item writing may be the often specific nature of the bias. Van Schilt-Mol (2007) identified item bias in educational tests (Cito tests) in Dutch primary schools using psychometric procedures. She then attempted to identify the source of the item bias, using a content analysis of the items and interviews with teachers and immigrant pupils. Based on this analysis, she changed the original items and administered the new version. The modified items showed little or no bias, indicating that she successfully identified and removed the bias source. Her study illustrates an effective, though laborious way to deal with bias. The source of the bias was often item specific (such as words or pictures that were not equally known in all cultural groups) and no general conclusions about how to avoid items could be drawn from her study.

Item bias has also been studied in personality and attitude measures. Although I do not know of any systematic comparison, the picture that emerges from the literature is one of great variability in numbers of biased items across instruments. There are numerous examples in which many or even a majority of the items turned out to be biased. If so many items are biased, serious validity issues have to be addressed, such as potential construct bias and adequate construct coverage in the remaining items. A few studies have examined the nature of item bias in personality questionnaires. Sheppard, Han, Colarelli, Dai, and King (2006) examined bias in the Hogan Personality Inventory in Caucasian and African Americans, who had applied for unskilled factory jobs. Although the group mean differences were trivial, more than a third of the items showed item bias. Items related to cautiousness tended to be potentially biased in favor of African Americans. Ryan, Horvath, Ployhart, Schmitt, and Slade (2000) were interested in determining sources of item bias global employee opinion surveys. Analyzing data from a 36-country study involving more than 50,000 employees, they related item bias statistics (derived from item response theory) to country characteristics. Hypotheses about specific item contents and Hofstede's (2001) dimensions were only partly confirmed; the authors found that more dissimilar countries showed more item bias. The positive relation between the size of global cultural differences and item bias may well generalize to other studies. Sandal et al. (in preparation) also found more bias between countries that are culturally

further apart. If this conclusion would hold across other studies, it would imply that a larger cultural distance between countries can be expected to be associated with more valid cross-cultural differences and more item bias. Bingenheimer, Raudenbush, Leventhal, and Brooks-Gunn (2005) studied bias in the Environmental Organization and Caregiver Warmth scales that were adapted from several versions of the HOME Inventory (Bradley, 1994; Bradley, Caldwell, Rock, Hamrick, & Harris, 1988). The scales are measures of parenting climate. There were about 4,000 Latino, African American, and European American parents living in Chicago that participated. Procedures based on item response theory were used to identify bias. Biased items were not thematically clustered.

1.3.2.4 Examples of Studies of Multiple Sources of Bias

Some studies have addressed *multiple sources of bias*. Thus, Hofer, Chasiotis, Friedlmeier, Busch, and Campos (2005) studied various forms of bias in a thematic apperception test, which is an implicit measure of power and affiliation motives. The instrument was administered in Cameroon, Costa Rica, and Germany. Construct bias in the coding of responses was addressed in discussions with local informants; the discussions pointed to the equivalence of coding rules. Method bias was addressed by examining the relation between test scores and background variables such as age and education. No strong evidence was found. Finally, using loglinear models, some items were found to be biased. As another example, Meiring, Van de Vijver, Rothmann, and Barrick (2005) studied construct, item, and method bias of cognitive and personality tests in a sample of 13,681 participants who had applied for entry-level police jobs in the South African Police Services. The sample consisted of Whites, Indians, Coloreds, and nine Black groups. The cognitive instruments produced very good construct equivalence, as often found in the literature (e.g., Berry, Poortinga, Segall, & Dasen, 2002; Van de Vijver, 1997); moreover, logistic regression procedures identified almost no item bias (given the huge sample size, effect size measures instead of statistical significance were used as criterion for deciding whether items were biased). The personality instrument (i.e., the 16 PFI Questionnaire that is an imported and widely used instrument in job selection in South Africa) showed more structural equivalence problems. Several scales of the personality questionnaire revealed construct bias in various ethnic groups. Using analysis of variance procedures, very

little item bias in the personality scales was observed. Method bias did not have any impact on the (small) size of the cross-cultural differences in the personality scales. In addition, several personality scales revealed low internal consistencies, notably in the Black groups. It was concluded that the cognitive tests were suitable as instruments for multicultural assessment, whereas bias and low internal consistencies limited the usefulness of the personality scales.

1.4 IDENTIFICATION OF BIAS IN STRUCTURAL EQUATION MODELING

There is a fair amount of convergence on how equivalence should be addressed in structural equation models. I mention here the often quoted classification by Vandenberg (2002; Vandenberg & Lance, 2000) that, if fully applied, has eight steps:

1. A global test of the equality of covariance matrices across groups.
2. A test of *configural invariance* (also labeled weak factorial invariance) in which the presence of the same pattern of fixed and free factor loadings is tested for each group.
3. A test of *metric invariance* (also labeled strong factorial invariance) in which factor loadings for identical items are tested to be invariant across groups.
4. A test of *scalar invariance* (also labeled strict invariance) in which identity of intercepts when identical items are regressed on the latent variables.
5. A test of invariance of unique variances across groups.
6. A test of invariance of factor variances across groups.
7. A test of invariance of factor covariances across groups.
8. A test of the null hypothesis of invariant factor means across groups. The latter is a test of cross-cultural differences in unobserved means.

The first test (the local test of invariance of covariance matrices) is infrequently used, presumably because researchers are typically more interested in modeling covariances than merely testing their cross-cultural

invariance and the observation that covariance matrices are not identical may not be informative about the nature of the difference. The most frequently reported invariance tests involve configural, metric, and scalar invariance (steps 2 through 4). The latter three types of invariance address relations between observed and latent variables. As these involve the measurement aspects of the model, they are also referred to as measurement invariance (or measurement equivalence). The last four types of invariance (steps 5 through 8) address characteristics of latent variables and their relations; therefore, they are referred to as structural invariance (or structural equivalence).

As indicated earlier, there is a confusing difference in the meaning of the term "structural equivalence," as employed in the cross-cultural psychology tradition, and "structural equivalence" (or structural invariance), as employed in the SEM tradition. Structural equivalence in the cross-cultural psychology tradition addresses the question of whether an instrument measures the same underlying construct(s) in different cultural groups and is usually examined in exploratory factor analyses. Identity of factors is taken as evidence in favor of structural equivalence, which then means that the structure of the underlying construct(s) is identical across groups. Structural equivalence in the structural equation tradition refers to identical variances and covariances of structural variables (latent factors) of the model. Whereas structural equivalence addresses links between observed and latent variables, structural invariance does not involve observed variables at all. Structural equivalence in the cross-cultural psychology tradition is much closer to what in the SEM tradition is between configural invariance and metric invariance (measurement equivalence) than to structural equivalence.

I now describe procedures that have been proposed in the structural equation modeling tradition to identify the three types of bias (construct, method, and item bias) as well as illustrations of the procedures; an overview of the procedures (and their problems) can be found in Table 1.1.

1.4.1 Construct Bias

1.4.1.1 Procedure

The structural equivalence tradition started with the question of how invariance of any parameter of a structural equation model can be tested. The aim of the procedures is to establish such invariance in a statistically

TABLE 1.1

Overview of Types of Bias and Structural Equation Modeling (SEM) Procedures to Identify These

Type of Bias	Definition	SEM Procedure for Identification	Problems
Construct	A construct differs across cultures, usually due to an incomplete overlap of construct-relevant behaviors.	Multigroup conformatory factor analysis, testing configural invariance (identity of patterning of loadings and factors).	Cognitive interviews and ethnographic information may be needed whether construct is adequately captured.
Method	Generic term for all sources of bias due to factors often described in the methods section of empirical papers. Three types of method bias have been defined, depending on whether the bias comes from the sample, administration, or instrument.	Confirmatory factor analysis or path analysis of models that evaluate the influence of method factors (e.g., by testing method factors).	Many studies do not collect data about method factors, which makes the testing of method factor impossible.
Item	Anomalies at the item level; an item is biased if respondents from different cultures with the same standing on the underlying construct (e.g., they are equally intelligent) do not have the same mean score on the item.	Multigroup confirmatory factor analysis, testing scalar invariance (testing identity of intercepts when identical items are regressed on the latent variables; assumes support for configural and metric equivalence).	Model of scalar equivalence, prerequisite for a test of items bias, may not be supported. Reasons for item bias may be unclear.

rigorous manner. The focus of the efforts has been on the comparability of previously tested data. The framework does not specify or prescribe how instruments have to be compiled to be suitable for cross-cultural comparisons; rather, the approach tests corollaries of the assumption that the instrument is adequate for comparative purposes. The procedure for addressing this question usually follows the steps described before, with

an emphasis on the establishment of configural, metric, and scalar invariance (weak, strong, and strict invariance).

1.4.1.2 Examples

Caprara, Barbaranelli, Bermúdez, Maslach, and Ruch (2000) tested the cross-cultural generalizability of the Big Five Questionnaire (BFQ), which is a measure of the Five Factor Model in large samples from Italy, Germany, Spain, and the United States. The authors used exploratory factor analysis, simultaneous component analysis (Kiers, 1990), and confirmatory factor analysis. The Italian, American, German, and Spanish versions of the BFQ showed factor structures that were comparable: "Because the pattern of relationships among the BFQ facet-scales is basically the same in the four different countries, different data analysis strategies converge in pointing to a substantial equivalence among the constructs that these scales are measuring" (p. 457). These findings support the universality of the five-factor model. At a more detailed level the analysis methods did not yield completely identical results. The confirmatory factor analysis picked up more sources of cross-cultural differences. The authors attribute the discrepancies to the larger sensitivity of confirmatory models.

Another example comes from the values domain. Like the previous study, it addresses relations between the (lack of) structural equivalence and country indicators. Another interesting aspect of the study is the use of multidimensional scaling where most studies use factor analysis. Fontaine, Poortinga, Delbeke, and Schwartz (2008) assessed the structural equivalence of the values domain, based on the Schwartz value theory, in a dataset from 38 countries, each represented by a student and a teacher sample. The authors found that the theoretically expected structure provided an excellent representation of the average value structure across samples, although sampling fluctuation causes smaller and larger deviations from this average structure. Furthermore, sampling fluctuation could not account for all these deviations. The closer inspection of the deviations shows that higher levels of societal development of a country were associated with a larger contrast between protection and growth values. Studies of structural equivalence in large-scale datasets open a new window on cross-cultural differences. There are no models of the emergence of constructs that accompany changes in a country, such as increases in the level

of affluence. The study of covariation between social developments and salience of psychological constructs is largely uncharted domain.

A third example from the values domain comes from Spini (2003), who examined the measurement equivalence of 10 value types from the Schwartz Value Survey in a sample of 3,859 students from 21 different countries. Acceptable levels of configural and metric equivalence were found for all values, except Hedonism. The hypothesis of scalar equivalence was rejected for all value types. Although the study by Fontaine et al. (2008) tested the universality of the global structure whereas Spini tested the equivalence of the separate scales, the two studies show remarkable resemblance in that structural equivalence was relatively well supported.

Arends-Tóth and Van de Vijver (2008) studied associations between well-being and family relationships among five cultural groups in the Netherlands (Dutch mainstreamers, and Turkish, Moroccan, Surinamese, and Antillean immigrants). Two aspects of relationships were studied: family values, which refer to obligations and beliefs about family relationships, and family ties that involve more behavior-related relational aspects. A SEM model was tested in which the two aspects of relationships predicted a latent factor, called well-being, which was measured by loneliness and general and mental health. Multisample models showed invariance of the regression weights of the two predictors and of the factor loadings of loneliness and health. Other model components showed some cross-cultural variation (correlations between the errors of the latent and outcome variables).

Van de Vijver (2002) examined the comparability of scores on tests of inductive reasoning in samples of 704 Zambian, 877 Turkish, and 632 Dutch pupils from the highest two grades of primary and the lowest two grades of secondary school. In addition to the two tests of inductive reasoning (employing figure and nonsense words as stimuli, respectively), three tests were administered that assessed cognitive components that are assumed to be important in inductive thinking (i.e., classification, rule generation, and rule testing). SEM was used to test the fit of a MIMIC model in which the three component tests predicted a latent factor, labeled inductive reasoning, which was measured by the two tests mentioned. Configural invariance was supported, metric equivalence invariance was partially supported, and tests of scalar equivalence showed a poor fit. It was concluded that comparability of test scores across these groups was problematic and that cross-cultural score differences were probably

influenced by auxiliary constructs such as test exposure. Finally, Davidov (2008) examined invariance of a 21-item instrument measuring human values of the European Social Survey that was administered in 25 countries. Multigroup confirmatory factor analysis did not support configural and metric invariance across these countries. Metric equivalence was only established after a reduction in the number of countries to 14 and of the original 10 latent factors to 7.

1.4.2 Method Bias

1.4.2.1 Procedure

The study of method bias in SEM is straightforward. Indicators of the source of method bias, which are typically viewed as confounding variables, can be introduced in a path model, which enables the statistical evaluation of their impact. Below examples of studies in response styles are given, but other examples can be easily envisaged, such as including years of schooling, socioeconomic status indicators, or interviewer characteristics. The problem with the study of method bias is usually not the statistical evaluation but the availability of pertinent data. For example, social desirability is often mentioned as a source of cross-cultural score differences but infrequently measured; only when such data are available, an evaluation of its impact can be carried out.

1.4.2.2 Examples

Various authors have addressed the evaluation of response sets, notably acquiescence and extremity scoring (e.g., Cheung & Rensvold, 2000; Mirowsky & Ross, 1991; Watson, 1992); yet, there are relatively few systematic SEM studies of method bias compared to the numerous studies on other types of bias. Billiet and McClendon (2000) worked with a balanced set of Likert items that measured ethnic threat and distrust in politics in a sample of Flemish respondents. The authors found a good fit for a model with three latent factors: two content factors (ethnic threat and distrust in politics that are negatively correlated) with positive and negative slopes according to the wording of the items, and one uncorrelated common style factor with all positive loadings. The style factor was identified as acquiescence, given that its correlation with the sum of agreements was

very high. Welkenhuysen-Gybels, Billiet, and Cambré (2003) applied a similar approach in a cross-cultural study.

1.4.3 Item Bias

1.4.3.1 Procedure

Item bias in SEM is closely associated with the test of scalar invariance. It is tested by examining invariance of intercepts when an item is regressed on its latent factor (fourth step in Vandenberg's procedure). The procedure is different from those described in the differential item functioning tradition (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993). Although it is impossible to capture the literally hundreds of procedures in this tradition that have been proposed, some basic ideas prevail. The most important is the relevance of comparing item statistics per score level. The latter are usually defined by splitting up a sample in subsamples of respondents with similar scores (such as splitting up the sample in low, medium, and high scorers). Corollaries of the assumption that equal sum scores on the (unidimensional) instrument reflect an equal standing on the latent trait are then tested. For example, the Mantel–Haenszel procedure tests, where the mean scores of persons with the same sum scores are identical across cultures (as they should be for an unbiased item). The SEM procedure tests whether the (linear) relation between observed and latent variable is identical across cultures (equal slopes and intercepts). From a theoretical point of view, the Mantel–Haenszel and SEM procedures are very different; for example, the Mantel–Haenszel procedure is based on a nonlinear relation between item score and latent trait whereas SEM employs a linear model. Also, both employ different ways to get access to the latent trait (through covariances in SEM and slicing up data in score levels in the Mantel–Haenszel procedure. Yet, from a practical point of view, the two procedures will often yield convergent results. It has been shown that using the Mantel–Haenszel is conceptually identical to assuming a Rasch model to apply to the scale and testing identity of item parameters across groups (Fischer, 1993). The nonlinear (though strictly monotonous) relation between item and latent construct score that is assumed in the Rasch model will often not differ much from the linear relation assumed by SEM. Convergence of results is therefore not surprising, in particular when items show a strong bias.

It is an attractive feature of SEM that biased items do not need to be eliminated from the instrument prior to the cross-cultural comparison (as are often done in analyses based on other statistical models). Biased items can be retained as culture-specific indicators. Partial measurement invariance allows for including both shared and nonshared items in cross-cultural comparisons. Scholderer, Grunert, and Brunsø (2005) describe a procedure for identifying intercept differences and correcting for these differences in the estimation of latent means.

1.4.3.2 Examples

Two types of procedures can be found in the literature that addresses item bias. In the first and most common type, item bias is part of a larger exercise to study equivalence and is tested after configural and metric equivalence have been established. The second kind of application adds information from background characteristics to determine to what extent these characteristics can help to identify bias.

De Beuckelaer, Lievens, and Swinnen (2007) provide an example of the first type of application. They tested the measurement equivalence of a global organizational survey that measures six work climate factors in 24 countries from West Europe, East Europe, North America, the Americas, Middle East, Africa, and the Asia-Pacific region; the sample comprised 31, 315 employees and survey consultants. The survey instrument showed configural and metric equivalence of the six-factor structure, but scalar equivalence was not supported. Many intercept differences of items were found; the authors argued that this absence was possibly a consequence of response styles. They split up the countries in regions with similar countries or with the same language. Within these more narrowly defined regions (e.g., Australia, Canada, United Kingdom, and the United States as the English-speaking region), scalar equivalence was found. A study by Prelow, Michaels, Reyes, Knight, and Barrera (2002) provides a second example. These authors tested the equivalence of the Children's Coping Strategies Checklist in a sample of 319 European American, African American, and Mexican American adolescents from low-income, inner-city families. The coping questionnaire consisted of two major styles, active coping and avoidant coping, each of which comprised different subscales. Equivalence was tested per subscale. Metric equivalence was strongly supported for all subscales of the coping questionnaire; yet,

intercept invariance was found in few cases. Most of the salient differences in intercept were found between the African American and Mexican American groups.

An example of the second type of item bias study has been described by Grayson, Mackinnon, Jorm, Creasey, and Broe (2000). These authors were interested in the question of whether physical disorders influence scores on the Center for Epidemiologic Studies Depression Scale (CES-D) among elderly, thereby leading to false-positives in assessment procedures. The authors recruited a sample of 506 participants aged 75 or older living in their community in Sydney, Australia. The fit of a MIMIC model was tested. The latent factor, labeled depression, was measured by the CES-D items; item bias was defined as the presence of significant direct effects of background characteristics on items (so, no cultural variation was involved). Various physical disorders (such as mobility disability and peripheral vascular disease) had a direct impact on particular item scores in addition to the indirect path through depression. The authors concluded that the CES-D score is "polluted with contributions unrelated to depression" (p. 279). The second example is due to Jones (2003), who assessed cognitive functioning among African American and European American older adults (>50 years) in Florida during a telephone interview. He also used a MIMIC model. Much item bias was found (operationalized here as differences in both measurement weights and intercepts of item parcels on a general underlying cognition factor). Moreover, the bias systematically favored the European American group. After correction for this bias, the size of the cross-cultural differences in scores was reduced by 60%. Moreover, various background characteristics had direct effects on item parcels, which were interpreted as evidence for item bias.

The two types of applications provide an important difference in perspective on item bias. The first approach only leads to straightforward findings if the null hypothesis of scalar equivalence is confirmed; if, as is often the case, no unambiguous support for scalar equivalence is found, it is often difficult to find reasons that are methodologically compelling for the lack of scalar equivalence. So, the conclusion can then be drawn that scalar equivalence is not supported and a close inspection of the deviant parameters will indicate those items that are responsible for the poor fit. However, such an observation usually does not suggest a substantive reason for the poor fit. The second approach starts from a more focused search for a specific antecedent of item bias. As a consequence, the

results of these studies are easier to interpret. This observation is in line with a common finding in item bias studies of educational and cognitive tests (e.g., Holland & Wainer, 1993): Without specific hypotheses about the sources of item bias, a content analysis of which items are biased and unbiased hardly ever leads to interpretable results as to the reasons for the bias.

The literature on equivalence testing is still scattered and is not yet ready for a full-fledged, meta-analysis of the links between characteristics of instruments, samples, and their cultures on the one hand, and levels of equivalence on the other hand; yet, it is already quite clear that studies of scalar equivalence often do not support the direct comparison of scores across countries. Findings working with SEM and findings based on other item bias techniques point in the same direction: Item bias is more pervasive than we may conveniently think and when adequately tested, scalar equivalence is often not supported. The widespread usage of analyses of (co)variance, *t*-tests, and other techniques that assume full score equivalence, is not based on adequate invariance testing. The main reason for not bothering about scalar invariance prior to comparing means across cultures is opportunistic: various studies have compared the size of cross-cultural differences before and after correction for item bias and most of these have found that item bias does not tend to favor a single group and that correction for item bias usually does not affect the size of cross-cultural differences (Van de Vijver, in press).

1.5 STATISTICAL MODELING AND BIAS: A SWOT ANALYSIS

After the description of a framework for bias and equivalence and a description of various examples in which the framework was employed, the stage is set for an evaluation of the contribution of SEM to the study of bias and equivalence. The evaluation takes the form of a SWOT analysis (strengths, weaknesses, opportunities, and threats).

The main *strength* of SEM is the systematic manner in which invariance can be tested. There is no other statistical theory that allows for such a fine-grained, flexible, and integrated analysis of equivalence. No other older approach combines these characteristics; for example, a combination of

exploratory factor analysis and item bias analysis could be used for examining the configural and scalar equivalence, respectively. However, the two kinds of procedures are conceptually unrelated. As a consequence, partial invariance is difficult to incorporate in such analyses. Furthermore, SEM has been instrumental in putting equivalence testing on the agenda of cross-cultural researchers and in stimulating the interest in cross-cultural studies.

The first *weakness* of equivalence testing using SEM is related to the large discrepancy between the advanced level of statistical theorizing behind the framework and the far from advanced level of available theories about cross-cultural similarities and differences. The level of sophistication of our conceptual models of cross-cultural differences is nowhere near the statistical sophistication available to test these differences. As a consequence, it is difficult to strike a balance between conceptual and statistical considerations in equivalence testing. The literature shows that it is tempting to use multigroup factor analysis in a mechanical manner by relying entirely on statistical, usually significance criteria to draw conclusions about levels of equivalence. An equivalence test using SEM can easily become synonymous with a demonstration that scores can be compared in a bias-free manner. In my view, there are two kinds of problems with these mechanical applications of equivalence tests. First, there are statistical problems with the interpretation of fit tests. Particularly in large-scale cross-cultural studies, the lack of convergence of information provided by the common fit statistics, combined with the absence of adequate Monte Carlo studies and experience with fit statistics in similar cases, can create problems in choosing the most adequate model. In these studies it is difficult to tease apart fit problems due to conceptually trivial sample particulars that do not challenge the interpretation of the model as being equivalent and fit problems due to misspecifications of the model that are conceptually consequential. Secondly, equivalence testing in SEM can easily become a tool that, possibly inadvertently, uses statistical sophistication to compensate for problems with the adequacy of instruments or samples. Thus, studies using convenience samples have problems of external validity, whatever the statistical sophistication used to deal with the data. Also, it is relatively common in cross-cultural survey research to employ short instruments. Such instruments may yield a poor rendering of the underlying construct and may capitalize on item specifics, particularly in a cross-cultural framework.

In addition to statistical problems, there is another and probably more salient problem of equivalence testing in a SEM framework: Sources of bias can be easily overlooked in standard equivalence tests based on confirmatory factor analysis, thereby reaching overly liberal conclusions about equivalence. Thus, construct inequivalence cannot be identified in deductive equivalence testing (i.e., testing in which only data from a target instrument are available, as is the case in confirmatory factor analysis). There is a tendency in the literature to apply closely translated questionnaires without adequately considering adaptation issues (Hambleton, Merenda, & Spielberger, 2005). Without extensive pretesting, the use of interviews to determine the accuracy of items, or the inclusion of additional instruments to check the validity of a target instrument, it is impossible to determine whether closely translated items are the best possible items in a specific culture. Culture-specific indicators of common constructs may have been missed. The focus on using identical instruments in many cultures may lead to finding superficial similarities between cultures, because the instrument compilation may have driven the study to an emphasis on similarities. The various sources of bias (construct, method, and items) cannot be investigated adequately if only data from the target instrument are available. Various sources of bias can be studied in SEM, but most applications start from a narrow definition of bias that capitalizes on confirmatory factor analysis without considering or having additional data to address bias. It should be noted that the problem of not considering all bias sources in cross-cultural studies is not an intrinsic characteristic of SEM, but a regrettable, self-imposed limitation in its use.

A first *opportunity* of equivalence testing using SEM is its scope to establish a closer link between the statistical modeling and inference levels. The discrepancy between the widespread usage of statistical techniques that compare mean scores across countries, such as analysis of variance, and the frequent observation in SEM procedures that conditions of scalar equivalents are not fulfilled defines a clear mission for SEM researchers. A second opportunity is related to the distinction between significance and relevance. It is quite clear that blind applications of significance testing often do not yield meaningful results; however, more work is needed to identify boundaries of practical significance. How much lack of fit can be tolerated before different substantive conclusions have to be drawn?

The main *threat* is that SEM procedures remain within the purview of SEM researchers. Usage of the procedures has not (yet?) become popular

among substantive researchers. There is a danger that SEM researchers keep on "preaching the gospel to the choir" by providing solutions to increasingly complex technical issues without linking questions from substantive researchers and determining how SEM can help to solve substantive problems and advance our theorizing.

1.6 CONCLUSIONS

Statistical procedures in the behavioral and social sciences are tools to improve research quality. This also holds for the role of SEM procedures in the study of equivalence and bias. In order to achieve a high quality, a combination of various types of expertise is needed in cross-cultural studies. SEM procedures can greatly contribute to the quality of cross-cultural studies, but more interaction between substantive and method researchers is needed to realize this potential. It is not a foregone conclusion that the potential of SEM procedures will materialize and that the threats of these procedures will not materialize. We need to appreciate that large-scale cross-cultural studies require many different types of expertise; it is unrealistic to assume that there are many researchers who have all the expertise required to conduct such studies. Substantive experts are needed with knowledge of the target construct, next to cultural experts with knowledge about construct in the target context, next to measurement experts who can convert substantive knowledge in adequate measurement procedures, next to statistical experts who can test bias and equivalence in a study. The strength of a chain is defined by the strength of the weakest link; this also holds for the quality of cross-cultural studies. SEM has great potential for cross-cultural studies, but it will be able to achieve this potential only in close interaction with expertise from various other domains.

REFERENCES

Aquilino, W. S. (1994). Interviewer mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly, 58,* 210–240.

Arends-Tóth, J. V., & Van de Vijver, F. J. R. (2008). Family relationships among immigrants and majority members in the Netherlands: The role of acculturation. *Applied Psychology: An International Review, 57,* 466–487.

Azhar, M. Z., & Varma, S. L. (2000). Mental illness and its treatment in Malaysia. In I. Al-Issa (Ed.), *Al-Junun: Mental illness in the Islamic world* (pp. 163–185). Madison, CT: International Universities Press.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., & Eysenck, H. J. (1998). The Eysenck personality questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences, 25,* 805–819.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology: Research and applications* (2nd ed.). New York: Cambridge University Press.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*, 608–628.

Bingenheimer, J. B., & Raudenbush, S. W., Leventhal T., & Brooks-Gunn, J. (2005) Measurement equivalence and differential item functioning in family psychology. *Journal of Family Psychology, 19,* 441–455.

Bradley, R. H. (1994). A factor analytic study of the infant-toddler and early childhood versions of the HOME Inventory administered to White, Black, and Hispanic American parents of children born preterm. *Child Development, 65,* 880–888.

Bradley, R. H., Caldwell, B. M., Rock, S. L., Hamrick H. M., & Harris, P. (1988). Home observation for measurement of the environment: Development of a home inventory for use with families having children 6 to 10 years old. *Contemporary Educational Psychology, 13,* 58–71.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Caprara, G. V., Barbaranelli, C., Bermúdez, J., Maslach, C., & Ruch, W. (2000). Multivariate methods for the comparison of factor structures. *Journal of Cross-Cultural Psychology, 31,* 437–464.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, 31, 160–186.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28,* 61–75.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods, 2,* 33–46.

De Beuckelaer, A., Lievens, F., & Swinnen, G. (2007). Measurement equivalence in the conduct of a global organizational survey across countries in six cultural regions. *Journal of Occupational and Organizational Psychology, 80,* 575–600.

De Jong, J. T. V. M., Komproe, I. V., Spinazzola, J., Van der Kolk, B. A., Van Ommeren, M. H., Marcopulos, F. (2005). DESNOS in three postconflict settings: Assessing cross-cultural construct equivalence. *Journal of Traumatic Stress, 18,* 13–21.

Emenogu, B. C. & Childs, R. A. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education, 28,* 128–146.

Fernández, A. L., & Marcopulos, B. A. (2008). A comparison of normative data for the trail making test from several countries: Equivalence of norms and considerations for interpretation. *Scandinavian Journal of Psychology, 49,* 239–246.

Fischer, G. H. (1993). Notes on the Mantel Haenszel procedure and another chi squared test for the assessment of DIF. *Methodika, 7,* 88–100.

Fischer, R., Fontaine, J. R. J., Van de Vijver, F. J. R., Van Hemert, D. A. (in press). What is style and what is bias in cross-cultural comparisons? An examination of acquiescent response styles in cross-cultural research. In A. Gari & K. Mylonas (Eds.), *Quod erat demonstrandum: From Herodotus' ethnographic journeys to cross-cultural research*. Athens: Atropos Editions.

AU:Can this be updated with publication information? Is the year 2007 correct here?

Fischer, R., & Mansell, A. (2007). *Levels of organizational commitment across cultures: A meta-analysis*. Manuscript submitted for publication.

Fontaine, J. R. J., Poortinga, Y. H., Delbeke, L., & Schwartz, S. H. (2008). Structural equivalence of the values domain across cultures: Distinguishing sampling fluctuations from meaningful variation. *Journal of Cross-Cultural Psychology, 39,* 345–365.

Formann, A. K., & Piswanger, K. (1979). *Wiener Matrizen-Test. Ein Rasch-skalierter sprachfreier Intelligenztest* [The Viennese Matrices Test. A Rasch-calibrated non-verbal intelligence test]. Weinheim, Germany: Beltz Test.

Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the center for epidemiologic studies depression scale: Effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology: Psychological Sciences, 55,* 273–282.

Hambleton, R. K., Merenda, P., & Spielberger C. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Hillsdale, NJ: Lawrence Erlbaum.

Harzing, A. (2006). Response styles in cross-national survey research: A 26-country study. *Journal of Cross Cultural Management, 6,* 243–266.

Ho, D. Y. F. (1996). Filial piety and its psychological consequences. In M. H. Bond (Ed.), *Handbook of Chinese psychology* (pp. 155–165). Hong Kong: Oxford University Press.

Hofer, J., Chasiotis, A., Friedlmeier, W., Busch, H., & Campos, D. (2005). The measurement of implicit motives in three cultures: Power and affiliation in Cameroon, Costa Rica, and Germany. *Journal of Cross-Cultural Psychology, 36,* 689–716.

Hofstede, G. (2001). *Culture's consequences. Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

AU:Please cite in text or remove from reference list.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton, NJ: Princeton University Press.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jones, R. N. (2003). Racial bias in the assessment of cognitive functioning of older adults. *Aging & Mental Health, 7,* 83–102.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70,* 631–639.

Kiers, H. A. L. (1990). *SCA: A program for simultaneous components analysis.* Groningen, the Netherlands: IEC ProGamma.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Lisse, the Netherlands: Swets & Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N., & Trewin, D. (1997). *Survey measurement and process quality*. New York: Wiley.

McCrae, R. R. (2002). Neo-PI-R data from 36 cultures: Further intercultural comparisons. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 105–125). New York: Kluwer Academic/Plenum Publishers.

Meiring, D., Van de Vijver, F. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item, and method bias of cognitive and personality tests in South Africa. *South African Journal of Industrial Psychology, 31,* 1–8.

Meyer, J. P., & Allen, N. J. (1991). A three-component conceptualization of organizational commitment. *Human Resource Management Review, 1,* 61–89.

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 × 2 index. *Social Psychology Quarterly* 54, 127–145.

Patel, V., Abas, M., Broadhead, J., Todd, C., & Reeler, A. (2001). Depression in developing countries: Lessons from Zimbabwe. *British Medical Journal, No. 322,* 482–484.

Piswanger, K. (1975). *Interkulturelle Vergleiche mit dem Matrizentest von Formann* [Cross-cultural comparisons with Formann's Matrices Test]. Unpublished doctoral dissertation, University of Vienna, Vienna.

Poortinga, Y. H. (1971). Cross-cultural comparison of maximum performance tests: Some methodological aspects and some experiments. *Psychologia Africana, Monograph Supplement, No. 6*.

Poortinga, Y. H. (1989). Equivalence of cross cultural data: An overview of basic issues. *International Journal of Psychology, 24,* 737–756.

Poortinga, Y. H., & Van der Flier, H. (1988). The meaning of item bias in ability tests. In S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context* (pp. 166–183). Cambridge: Cambridge University Press.

Prelow, H. M., Michaels, M. L**.,** Reyes, L., Knight, G. P., & Barrera, M. (2002). Measuring coping in low income European American, African American, and Mexican American adolescents: An examination of measurement equivalence. *Anxiety, Stress, and Coping, 15,* 135–147.

Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology, 53,* 541–562.

Sandal, G. M., Van de Vijver, F. J. R., Bye, H. H., Sam, D. L., Amponsah, B., Cakar, N., . . . Kosic, A. (in preparation). *Intended Self-Presentation Tactics in Job Interviews: A 10-Country Study*.

Scholderer, J., Grunert, K. G., & Brunsø, K. (2005). A procedure for eliminating additive bias from cross-cultural survey data. *Journal of Business Research, 58,* 72–78.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317–375.

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan personality inventory. *Assessment, 13,* 442–453.

Sireci, S. (in press). Evaluating test and survey items for bias across languages and cultures. In D. M. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology*. Cambridge: Cambridge University Press.

Spini, D. (2003). Measurement equivalence of 10 value types from the Schwartz value survey across 21 countries. *Journal of Cross-Cultural Psychology, 34,* 3–23.

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

Suzuki, K., Takei, N., Kawai, M., Minabe, Y., & Mori, N. (2003). Is Taijin Kyofusho a culture-bound syndrome? *American Journal of Psychiatry, 160,* 1358.

Tanaka-Matsumi, J., & Draguns, J. G. (1997). Culture and psychotherapy. In J. W. Berry, M. H. Segall, & C. Kagitcibasi (Eds.), *Handbook of cross-cultural psychology* (Vol. 3, pp. 449–491). Needham Heights, MA: Allyn and Bacon.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139–158.

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, 2,* 4–69.

Van de Vijver, F. J. R. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28,* 678–709.

Van de Vijver, F. J. R. (2002). Inductive reasoning in Zambia, Turkey, and the Netherlands: Establishing cross-cultural equivalence. *Intelligence, 30,* 313–351.

Van de Vijver, F. J. R. (in press). Bias and real differences in cross-cultural differences: Neither friends nor foes. In S. M. Breugelmans, A. Chasiotis, & F. J. R. van de Vijver (Eds.), *Fundamental questions in cross-cultural psychology*. Cambridge: Cambridge University Press.

Van de Vijver F. J. R., & Fischer, R. (2009). Improving methodological robustness in cross-cultural organizational research. In R. S. Bhagat & R. M. Steers (Eds.), *Handbook of culture, organizations, and work* (pp. 491–517). Cambridge, NY: Cambridge University Press.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

Van de Vijver, F. J. R., & Leung, K. (in press). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. M. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology*. Cambridge: Cambridge University Press.

AU:Please cite in text or remove from reference list.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Dordrecht: Kluwer.

AU:Please cite in text or remove from reference list.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology, 33,* 141–156.

Van Hemert, D. A., Van de Vijver, F. J. R., Poortinga, Y. H., & Georgas, J. (2002). Structural and functional equivalence of the Eysenck personality questionnaire within and between countries. *Personality and Individual Differences, 33,* 1229–1249.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*, 346–360.

Van Leest, P. F. (1997). Bias and equivalence research in the Netherlands. *European Review of Applied Psychology, 47,* 319–329.

Van Schilt-Mol, T. M. M. L. (2007). *Differential Item Functioning en itembias in de Cito-Eindtoets Basisonderwijs*. Amsterdam: Aksant.

Wasti, S. A. (2002). Affective and continuance commitment to the organization: Test of an integrated model in the Turkish context. *International Journal of Intercultural Relations, 26,* 525–550.

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness. *Sociological Methods Research, 21,* 52–88.

Welkenhuysen-Gybels, J., Billiet, J., & Cambre, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology, 34,* 702–722.