Bias and real differences in cross-cultural differences

van de Vijver, F.J.R.

Published in:
Fundamental questions in cross-cultural psychology

Publication date:
2011

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

10

Bias and real differences in cross-cultural differences: neither friends nor foes

FONS J. R. VAN DE VIJVER

> Though the difficulty of establishing comparability is widely acknowledged, the challenge is more often ignored than met.
>
> (Smith, 2003, p. 69)

The chapter starts from the above quotation by Tom Smith, who complained in a book on cross-cultural survey methods that comparability issues in cross-cultural surveys are more often mentioned than addressed. In a similar vein, Bollen, Entwistle and Alderson (1993) found in a meta-analysis of macrocomparative studies that equivalence is infrequently addressed. The situation in cross-cultural psychology is not much different; there is a widely acknowledged, shared awareness of potential pitfalls of direct cross-cultural score comparisons, but the sensitivity for the issue is insufficiently accompanied by tests of instrument adequacy in a specific study. It was argued in the first chapter by the editors that method issues are at the core of cross-cultural psychology. This chapter discusses one specific method issue, namely possible sources of bias in cross-cultural studies and the ramifications of bias for the cross-cultural comparability of scores. The editors mentioned in their chapter that the methodological problems of cross-cultural studies were already described seventy years ago and that many empirical researchers, methodologists and psychometricians have tried to tackle these problems. The question is addressed here to what extent we have advanced in this field. The present chapter deals with the question of how we should evaluate the current situation with regard to the study of bias. It is argued that the separation of valid differences and bias is counterproductive and that it is more productive to view bias as examples of culture-specific aspects of a measure. These culture-specific elements can be due to a wide variety of reasons, such as cross-cultural differences in construct definitions, poor translations or other aspects of a measure. The widely shared assumption according to which all bias has to be eliminated before cross-cultural comparisons can be made is questioned.

The interest in methodological aspects of cross-cultural comparisons is quite old, as has been repeatedly described by Jahoda (1982, this volume). The advent of modern cross-cultural psychology has given further impetus.

The first, systematic approach to measurement issues in cross-cultural psychology has probably been developed by the South African psychologist Simon Biesheuvel (1943, 1958). His work has remained relatively unnoticed in mainstream psychology; the relatively isolated position of South African psychology in the Apartheid era has probably contributed to this. Poortinga (1971) was among the first who linked comparability, which was the more common term in those days, to statistical modelling. Furthermore, he used the notion of hierarchically nested levels of comparability. In the following decades major developments in statistics, such as the development of structural equation modelling and test hierarchy, have enlarged the tool kit of cross-cultural psychologists considerably. Moreover, various concepts have been introduced to match the great level of detail that can be obtained by equivalence tests nowadays. Still, the original ideas of linking comparability to statistical modelling and establishing hierarchies of comparability have never changed; most developments were statistical refinements of the original ideas.

This chapter deals with the discrepancy between recommended and actual practices in cross-cultural psychology vis-à-vis equivalence testing. Have these tests become too difficult and impractical so that they are only used by methodological diehards and do not solve but create problems? Or do the tests not address a practical need? It is argued in the present chapter that the computational complexity is only a contributing factor. Another factor is the dichotomisation of valid (genuine) cross-cultural differences and bias. Current equivalence analyses start from the assumption that cross-cultural differences are either valid or due to bias. It is possibly more realistic to abandon this dichotomy and to start working from the assumption that even at the smallest level, such as items, cross-cultural differences often reflect both real differences and bias. The first part of the chapter provides a short overview of bias and equivalence techniques. The second part presents examples of the studies that address each of the types of bias described in the first part. Examples in the chapter are selected to represent a wide variety of psychology fields so as to illustrate the relevance and applicability of systematic sources of variance that derive from other sources than the construct intended to be measured. The third part describes the contours of an approach in which bias and valid differences are not treated as antithetical but as complementary sources of cross-cultural differences. Conclusions are described in the fourth part.

Bias and equivalence: taxonomy

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 and Poortinga, 1991; van de Vijver and Leung, 1997). There is *construct bias* if a construct that is measured by a test differs across cultures, usually due to an incomplete overlap of construct-relevant behaviours. 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 conception, which includes the expectation that children should assume the role of caretaker of elderly parents, is broader than the corresponding Western notion.

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 can be distinguished, 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 which 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 were more likely to display positive attitudes to an interviewer (e.g., Aquilino, 1994). A last example of administration bias can be found in communication problems between the respondent and the tester/interviewer. Instrument bias is a final source of bias in cognitive tests that refers to 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.

Item bias or *differential item functioning* refers to anomalies at the item level (Camilli and Shepard, 1994; Holland and Wainer, 1993). According to a definition that is widely used in education sciences and psychology, an item is biased if respondents with the same standing on the underlying construct (e.g., they are equally *intelligent*) do not have the same mean score on the item because of different cultural origins. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Camilli and Shepard, 1994; Holland and Wainer, 1993; van de Vijver and Leung, 1997, 2009; Sireci, 2009).

Although 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 administered to pupils in all EU countries asks 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; this will hold even for pupils from the two countries with the same level of knowledge in geography. The item is biased because it favours one cultural group across all test score levels.

Equivalence

Bias has implications for the comparability of scores (e.g., Poortinga, 1989). Building on the bias taxonomy presented above, four hierarchically nested types of equivalence are described below: construct, structural or functional, metric (or measurement unit) and scalar (or full score) equivalence.

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 opt for a strong emic, relativistic viewpoint, which argues that psychological constructs are inextricably tied up to their natural context and cannot be studied outside this 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 occurs in Asian countries, such as Indonesia and Malaysia. It is characterised by a brief period of violent aggressive behaviour among men. The period is often preceded by an insult and the patient shows persecutory ideas and automatic behaviours. After the period, the patient is usually exhausted and has no recollection of the event (Azhar and Varma, 2000). Violent aggressive behaviour 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 all culture-bound syndromes. The case of the Japanese Taijin Kyofusho is another example (Suzuki *et al.*, 2003; Tanaka-Matsumi and Draguns, 1997). Taijin Kyofusho is characterised 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 that do not co-occur in other cultures.

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 is then the consequence of a 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.

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. In operational terms, this condition requires identity of underlying dimensions (factors) in all groups, namely that the instrument shows the same factor structure in all groups. Structural equivalence has been examined for various cognitive tests (Jensen, 1980), Eysenck's Personality Questionnaire (Barrett *et al.*, 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. A questionnaire that measures, say, openness to new cultures shows functional equivalence in a study if it measures the same psychological constructs in each culture, as manifested in a similar pattern of convergent and divergent validity (i.e., non-zero 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-technical. With advances in statistical modelling (notably path analysis as part of structural equation modelling), 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 easily lead to an endless regression in which each scale used for validation has itself to be validated and scrutinized for equivalence.

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). Measurement unit equivalence applies when a source of bias shifts the scores of different cultural groups differentially, 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. When the relative contribution of both bias sources cannot be estimated, the interpretation of group comparisons of mean scores remains ambiguous.

Scalar or full score equivalence. Only in the case of scalar (or full-score) equivalence can direct cross-cultural comparisons be made; this is the only type of equivalence that allows for the conclusion that average scores obtained

in two cultures are different or equal. Scalar equivalence assumes an identical interval or ratio scale across groups.

Bias and equivalence: assessment and applications

Procedures. There are procedures in which data obtained with an instrument are sufficient to address bias and equivalence; there are also procedures that rely on data obtained with additional instruments to assess bias and equivalence in the target instrument. The latter procedures can be called open, inductive or exploratory, whereas the former can be called closed, deductive or hypothesis testing.

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 determine which attitudes and behaviours 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 data about 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 nearly always 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.

The assessment of structural equivalence also employs closed procedures. Correlations, covariances and distance measures between items or subtests can be 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 favour of structural equivalence. The absence of structural equivalence is interpreted as evidence in favour of construct inequivalence. Structural equivalence techniques, examples of 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. Testing metric and scalar equivalence

is also based on closed procedures. Structural equation modelling is often used to assess relations between items or subtests and their underlying constructs. It can be concluded that open and closed procedures use different methods to address equivalence. Closed procedures can be employed with any instrument with multiple items or subscales. Whether such a closed procedure can address all relevant equivalence issues depends on various factors. Collecting additional data is more important 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 there is a more pressing need to develop a culturally appropriate measure (possibly with culturally specific items or scales).

Examples. An interesting study of *construct bias* has been reported by Patel *et al.* (2001). These authors were interested in the question of 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 conceptualisation of depression is only partly overlapping with Western theories models. As a consequence, Western instruments will have a limited suitability, particularly with regard to the aetiology of the syndrome.

As another example, it has been argued that an organisational commitment also contains both shared and culture-specific components. Most Western research is based on a three-componential model (e.g., Meyer and Allen, 1991; see van de Vijver and Fischer, 2009) that differentiates between affective, continuance and normative commitment. Affective commitment is the emotional attachment to organisations, the desire to belong to the organisation and identification with the organisational norms, values and goals. Normative commitment is considered as a feeling of obligation to remain with the organisation, capturing normative pressures and perceived obligations by important others. Finally, continuance commitment refers to the costs associated with leaving the organisation and the perceived need to stay. Wasti (2002) argued that the concept of continuance commitment does not cover all relevant aspects in collectivistic contexts like Turkey. Loyalty and trust are strongly associated with paternalistic management practices. Employers

are more likely to give trusted jobs to family members or friends, involving these individuals into relationships of dependency and obligation. 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 countries. These findings suggest that the three components become more independent with increasing economic affluence.

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' (pp. 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 of the 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 the pattern and extent of support among family members in five cultural groups in the Netherlands (Dutch mainstreamers, Turkish-, Moroccan-, Surinamese-, and Antillean-Dutch). The authors made a distinction between provided support (i.e., what you give to other family members) and received support (i.e., what you receive from them). 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 (before correction) to .03 (after correction) for support and from .07 to .03 for contact.

Response styles and social desirability (which is usually not viewed as a response style, but involves a closely related concept) are often viewed as sources method bias; this is also done here. There is an ongoing debate whether these styles should not be treated as sources of substantive cross-cultural differences, as they are influenced by cross-cultural differences in conformity or communication styles (Johnson and van de Vijver, 2003). The styles reflect systematic variance that cannot be accounted for by the target construct of an attitude or personality questionnaire; therefore, they are viewed as bias source here. The study of response styles 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 North-western countries in surveys on consumer research. They interpreted these differences in terms of the individualism versus collectivism dimension. In a meta-analysis across forty-one countries, Fischer *et al.* (2009) calculated acquiescence scores for various scales in the personality, social-psychological and organisational domains. A small but significant percentage (3.1 per cent) 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 twenty-six 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, extroversion (at 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. Findings on social desirability also point to the presence of systematic cross-cultural differences. More affluent countries tend to show lower scores on social desirability (van Hemert *et al.*, 2002).

Instrument bias is another common source of method bias in cognitive tests. An example can be found in Piswanger's (1975) application of the Viennese Matrices Test (Formann and Piswanger 1979). A Raver-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.

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 which should be eliminated. As a consequence, items that are presumably biased are eliminated from the cross-cultural comparison. It is only after all biased items have been eliminated that cross-cultural differences can be adequately evaluated. 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 and 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 ten countries, it was found that dress code yielded biased items. 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) (Sandal *et al.*, in preparation). Clearly, these items provide important information about self-presentation in these countries, which cannot be dismissed as item bias that should be eliminated.

The forty years of item bias research have not led to aggregated insights as to which kind of items tend to be biased. In fact, one of the complaints has been the lack of accumulated insights. Educational testing has been an important domain of application of item bias; many techniques and applications have been presented in the *Journal of Educational Measurement*. Linn (1993) has reviewed the field with a view to integrating its findings. He came to the sobering conclusion that no general conclusions can be drawn about which item characteristics are associated with 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 building up accumulated knowledge about item writing may be the often specific nature of item 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 items and administered the modified version in new groups. The new items showed little or no bias, indicating that the bias source was successfully identified and removed. 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. Her study illustrates an effective, though laborious, way to eliminate bias.

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 *et al.* (2006) examined bias in the Hogan Personality Inventory across ethnic groups (Caucasian and African American) 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 favour of African Americans. Ryan *et al.* (2000)

were interested in determining sources of item bias global employee opinion surveys. Analysing data from a thirty-six-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; yet, 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 generalise 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 both more valid cross-cultural differences and more item bias.

Many studies of bias focus on the identification of a single source of bias (such as items). However, such a restriction is not obvious from a conceptual point of view. The bias taxonomy that is presented here does not treat bias sources as mutually exclusive. For example, it is possible that method bias (e.g., social desirability) co-occurs with item bias. As a consequence, it is important to examine multiple sources of bias. A few of such studies have been published. Thus, Hofer *et al.* (2005) studied various forms of bias in the Thematic Apperception Test, which is a measure of implicit motives (power and affiliation). 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 for the presence of method bias was found. Finally, item bias analysis was addressed using loglinear models. Some items were found to be biased. Their study clearly demonstrates that the validity of cross-cultural comparisons can be greatly enhanced by addressing various forms of bias. As another example, Meiring *et al.* (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, coloureds and nine black groups. The cognitive instruments produced very good construct equivalence, as often found in the literature (e.g., Berry *et al.*, 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, which 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.

The above studies attempted to identify one or more sources of bias and to draw conclusions about equivalence by addressing bias sources. In addition, there are many studies that address equivalence more directly. I give examples here of studies of the various levels of equivalence described before. There are few studies that are aimed at demonstrating *construct inequivalence*. However, various studies that addressed structural equivalence found construct inequivalence by showing that the underlying constructs were not (entirely) comparable. For example, De Jong *et al.* (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 victimisation, notably the consequences of war, genocide, persecution, torture and terrorism. The interview covers six clusters, each with two to six 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 multi-sample 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 Posttraumatic Stress Disorder that are included in DSM-IV. 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 multi-step interdisciplinary method that accommodates universal chronic sequelae of

extreme stress and accommodates culture-specific symptoms across a variety of cultures. The procedure illustrates well that constructs with only a partial overlap across cultures do not create a problem for cross-cultural comparison; these constructs just 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), that assumes that the core of psychological constructs tends to be invariant across cultures but manifestations may take culture-specific forms.

Many studies have addressed *structural equivalence*; most implications dealt with comparisons of a relatively small number of cultures. However, some applications involving a large number of countries have been reported. Probably the best-known examples come from the domain of personality. McCrae and Allik (2002) addressed the universality of the five-factor model of personality and presented impressive evidence for the universality of the five factors. Another example comes from studies of the Eysenck Personality Questionnaire (Barrett *et al.*, 1998), in which it is found that the three factors that constitute personality in Eysenck's model (extroversion, neuroticism and psychoticism) generalised fairly well across thirty-eight countries. I present a few illustrations of cross-cultural studies in which structural equivalence was examined in an interesting manner. Caprara *et al.* (2000) tested the cross-cultural generalisability 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 addressed equivalence using 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 are in line with the universality of the five-factor model just mentioned. At a more detailed level the analysis method did not yield completely identical results. The exploratory factor analysis and simultaneous component analysis are closely related whereas confirmatory factor analysis is based on a more hypothesis-testing approach. The latter pointed to relatively small cross-cultural differences in the factor structure. The authors attribute the discrepancies to the larger sensitivity of confirmatory models.

Van de Vijver and Poortinga (2002), analysing data from the 1990–1 World Values Survey (Inglehart, 1997), examined the structural equivalence of the post-materialism scale across thirty-nine countries. Postmaterialists tend to emphasise self-expression and quality of life as ulterior attitudes, whereas materialists emphasise economic and physical security above all (Inglehart, 1997, p. 4). It is Inglehart's thesis that with the increase of national affluence,

there is a shift from materialist to postmaterialist attitudes. The inventory comprised of nine items, such as 'Seeing that people have more to say about how things are done at their jobs and in their communities' and 'Trying to make our cities and countryside more beautiful'. A pairwise comparison would not be practical with so many countries in the data set. Therefore, it was decided to compute a pooled solution, based on the data of all countries, to which the factor structure of each country was compared. A good fit was found for nearly all countries; yet, it was found that the internal consistency of the scale increased with the level of affluence of the country. So, Inglehart's thesis provides only part of the picture; in addition to endorsement, the salience of the concept increases with growing affluence (and its wide range of accompanying changes such as increases in educational and healthcare expenditures).

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 multi-dimensional scaling where most studies use factor analysis. Fontaine *et al.* (2008) assessed the structural equivalence of the values domain, based on the Schwartz value theory, in a data set from thirty-eight 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 show that higher levels of societal development of a country were associated with a larger contrast between protection and growth values.

Spini (2003) examined the measurement equivalence of ten value types from the Schwartz Value Survey in a sample of 3,859 students from twenty-one different countries. Using nested multi-group confirmatory factor analyses, the author investigated the three most commonly tested levels of invariance: configural (akin to structural equivalence in exploratory factor analysis), metric and scalar invariance (van de Vijver and Leung, 1997; Vandenberg and Lance, 2000). Acceptable levels of configural and metric equivalence were found for all values, except Hedonism. The hypotheses of scalar and reliability equivalence were 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, which involve more behaviour-related relational aspects. Structural equation modelling was used in which the two aspects of relationships predicted a latent factor, called well-being, that was measured by loneliness and general and mental health. Multi-sample 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). The metric invariance of the model confirmed that relations among the target constructs are invariant across groups. Cross-cultural differences in family relationships (the cultural differences in mean scores between immigrants and majority members were larger for family values than for family ties) have to be interpreted against a backdrop of similar associations of these relationships with well-being.

As a final example, 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 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). Structural equation modelling was used to test the fit of a model in which the three component tests predicted a latent factor, labelled inductive reasoning, which was measured by the two tests mentioned. Configural invariance was supported, metric equivalence invariance was partially supported, but 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.

Several conclusions can be drawn from this section. Firstly, the few examples of tests of scalar equivalence described here could be complemented by many other examples which show that this type of equivalence is more often assumed than observed and that reported cross-cultural score comparisons are often based on insufficient methodological justification to warrant score comparisons. The question to what extent different conclusions would have been reached if comparisons would have been restricted to methodologically justifiable comparisons is not easy to answer. Secondly, studies of structural equivalence in large-scale datasets open a new window on cross-cultural differences. There are no models of the emergence of psychological constructs that accompany changes in a country, such as increases in postmodernity with increasing levels of affluence. The study of covariation between social developments and salience of psychological constructs is an uncharted, though relevant, domain for cross-cultural psychologists.

Ways forward

I see two areas of development in the study of bias and equivalence. The first involves overcoming the lack of balance in current approaches to bias and equivalence. The preponderance of studies of item bias and the relative absence of studies of other types of bias are both remarkable and regrettable. The lack of balance may suggest that items are the major source of inequivalence in cross-cultural studies. From a conceptual point of view, however, there is little reason for such a preference. It is difficult to see why sources of inequivalence are more likely at item than at construct or method level. Sample bias, response styles and partial construct non-overlap are sources of inequivalence with a pervasive influence on cross-cultural differences and may indeed be more consequential for conclusions drawn from cross-cultural studies if unchecked. Furthermore, removal of item bias often does not have a major impact on the size of the cross-cultural score differences observed (e.g., Meiring *et al.*, 2005). The field of cross-cultural psychology (and cross-cultural assessment in particular) is better served by a more balanced treatment of various bias sources. The study of bias sources should start from an analysis of potential threats of cross-cultural comparability rather than considerations of convenience or fashion. I mention a few ways to achieve a more integrated analysis of bias sources. An effective way to deal with method bias is the assessment of presumably relevant participant and contextual variables, such as education or score on individualism–collectivism (to deal with the often used assumption that all members of a culture have the same standing on these constructs). Also, the structural equivalence of scales should be investigated prior to the item bias analysis to establish that the scales are unidimensional in measuring the same dimension in each culture. Finally, cross-cultural psychology could learn from survey research where cognitive pretesting is used as an effective way to study construct bias.

A second line of development is more conceptual and involves the way in which we view bias (Poortinga and van der Flier, 1988). It is a fairly common finding that cultures that are further apart show more sources of bias. Bias is clearly a function of cultural distance. However, the size of observed cross-cultural differences in psychological test and inventories also tends to increase with cultural distance. So, we are left with what seems to constitute a paradox: from a conceptual perspective, bias and valid cross-cultural differences are unrelated and all our statistical models are based on their independence, but the conceptual independence is not accompanied by empirical independence. The separation between bias and valid differences which underlies so much work in cross-cultural psychology is fruitful from a statistical perspective, but counterproductive from a conceptual perspective. It is important to examine the ramifications of this empirical association between bias and valid differences.

My proposal is not to abandon bias testing, but to integrate this testing more in the analysis of cross-cultural differences and to make a distinction between two sources of bias. The first refers to errors in cross-cultural studies, such as bad translations or the application of items that do not have any meaning in a specific cultural context. A properly conducted study will show few of such errors. The second refers to all other systematic cross-cultural differences that cannot be interpreted in terms of the target construct and require other sources of validation, such as additional measures and in-depth interviews with participants. These latter differences are better viewed as culture-specific elements of the instruments used. Our tendency to remove these elements from the cross-cultural comparisons and to treat these as error leads to an overestimation of the common components and a neglect of culture-specific components of a construct (as we typically discuss only results based on the equivalent parts of an instrument). The components of the instrument that do not show this bias point to shared components of the target construct. We should replace the dichotomy between cross-cultural differences that are either valid or due to bias by four concepts. The first type of differences (differences based on shared components) refers to what has traditionally been viewed as valid cross-cultural differences. Tests of these differences are based on finding scalar equivalence interval or ratio scales. The major extension that is proposed here amounts to further splitting a bias in three types. The first type of bias involves all measurement artefacts, such as bad translations, floor and ceiling effects and interviewer effects; in sum, these refer to cross-cultural differences that cannot be interpreted in terms of the target construct and could be avoided in a replication by essentially technical changes in the instrument. It is important to note that the measure and not the target construct is here the source of variation. The second involves bias differences that can be interpreted in terms of relevant cultural differences; for example, items in a depression scale that involve psychological symptoms may show lower factor loadings in countries in which depression is primarily expressed through semantic symptoms. Although the items are biased, the interpretation of differences in loadings is crucial for the understanding of cross-cultural differences in depression. Conclusions of the study would be inadequate if these differences would not be taken into account. The third involves bias differences that cannot (yet) be interpreted adequately and that are unaccounted for. From a conceptual point of view, this is the most problematic type of bias, as it distorts cross-cultural comparisons for no obvious reasons. In the item bias tradition, where it often turns out to be difficult to understand the nature of the bias, the last category is heavily represented.

Bias and real cross-cultural differences have traditionally been treated as enemies. The dichotomy is based on the reasoning that cross-cultural differences can only be real if these are not shown to be biased. The present chapter has attempted to highlight the sterile and counterproductive nature of this

dichotomy. We should avoid starting a semantic discussion on the distinction between bias and real differences in such a case, and rather focus on the *need to account for the cross-cultural score differences* on the basis of other explanatory variables than the target construct (Poortinga *et al.*, 1987). A search for explanations of cross-cultural differences is usually based on the statistical evaluation of explanatory variables (e.g., these variables are used as covariates in an analysis of covariance or as predictors in a multiple regression analysis). Alternatively, cross-cultural differences are often interpreted post hoc if a statistical evaluation is not possible or unexpected differences have arisen. In such a search, the distinction between valid cross-cultural differences and bias is not always relevant. The main question is whether the explanatory variable is helpful in understanding cross-cultural differences, no matter whether these are based on valid differences or bias. The idea of unpackaging culture (see Bond's chapter in this volume) rests on the idea that observed cross-cultural differences are often not easy to interpret and require validation for an adequate interpretation. The explanatory variables provide such a validation. Suppose that we have administered a reading achievement test to groups of school-going children in two countries. Possible cross-cultural differences in age or motivation would probably be treated as sources of bias. The underlying idea is that if these groups were matched on age and motivation, the cross-cultural differences would change (in practice, the differences would presumably become smaller after correction). Possible cross-cultural differences in reading instruction could be viewed as an explanation of valid differences. So, a statistical correction for the quality or quantity of reading instruction would provide an explanation of (part of the) valid differences. From a conceptual point of view, both kinds of explanatory variables may well be complementary in that each provides a part of the explanation of the observed cross-cultural reading differences. The distinction between bias and valid differences may often be more theoretically than practically relevant in the explanation of cross-cultural differences.

We have advanced statistical techniques to identify multiple sources of bias. Our experiences with bias and equivalence analyses indicate that decisions about bias that are entirely based on statistical grounds can lead to various debatable inferences, such as the elimination of construct-relevant cross-cultural differences and the elimination of items that have hardly any impact on the size of the cross-cultural differences observed. A combination of statistical and cultural expertise is needed to conduct and evaluate equivalence analyses. Statistical tools can go a long way to identify bias, but do not provide valuable information as to the reasons of the bias; the latter requires cultural expertise. In recent years there has been a strong increase in the interest in mixed methods (Tashakkori and Teddlie, 2003); these involve the combined use of qualitative and quantitative methods. If these methods are properly used, conclusions can be drawn beyond the reach of mono-method studies. Mixed

methods could play an important role in reducing the number of biased stimuli that cannot be accounted for. For example, detailed analyses of items that show bias in the population of interest can help to gain insight in its nature; qualitative methods can provide important tools where the end of information that can be gained by quantitative methods has been reached.

Conclusion

Why is the study of equivalence so important if, as argued here, the relevance of the distinction between bias and valid differences is downplayed from a conceptual perspective? The crux of the argument is that we will not learn much from rigidly applying the dichotomy between bias and real differences. Finding cross-cultural differences in scores is the starting point, not the endpoint of cross-cultural studies. We can learn much by focusing on the interpretation of cross-cultural differences and on validating these differences. The study of bias and equivalence can help us to (dis)confirm interpretations of score differences. We reach more valid conclusions if we can be sure that the cross-cultural score differences we observe are not due to measurement anomalies; the study of bias and equivalence increases our insight in the quality of the measures we have used; finally, bias analyses are effective tools to rule out alternative explanations. Bias and equivalence analyses can help us to unpackage cross-cultural score differences (Poortinga and Van de Vijver, 1987) and 'to peel the onion called culture' (Poortinga *et al.*, 1987). However, the present chapter has illustrated how conceptual and statistical analyses of bias have developed along their own lines; an ever-continuing refinement of psychometric procedures to identify bias is unlikely to advance the field considerably. Rather, we should seek to re-establish the interactions between substantive and method researchers. The field will only advance if theory and method are combined. Blind applications of statistical tools on conceptually poor instruments and inadequate use of statistical tools on conceptually sophisticated instruments will both lead to results that are not convincing. A more judicious, theory-driven use of bias and equivalence techniques will help us to better understand what is shared across cultures and what is unique for cultures.

References

Aquilino, W. S. (1994). Interviewer mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly, 58*, 210–40.

Arends-Tóth, J. V., and 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*, 466–87.

Azhar, M. Z., and 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. 163–85). Madison, CT: International Universities Press.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., and 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*, 805–19.

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

Biesheuvel, S. (1943). *African intelligence*. Johannesburg: South African Institute of Race Relations.

(1958). Objectives and methods of African psychological research. *Journal of Social Psychology*, *47*, 161–8.

Bollen K. A., Entwisle, B., and Alderson, A. S. (1993). Macrocomparative research methods. *Annual Review of Sociology*, *19*, 321–51.

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

Caprara, G. V., Barbaranelli, C., Bermudez, J., Maslach, C., and Ruch, W. (2000). Multivariate methods for the comparison of factor structures. *Journal of Cross-Cultural Psychology*, *31*, 437–64.

De Jong, J. T. V. M., Komproe, I. V., Spinazzola, J., Van der Kolk, B. A., Van Ommeren, M. H., and Marcopulos, F. (2008). DESNOS in three postconflict settings: Assessing cross-cultural construct equivalence. *Journal of Traumatic Stress*, *18*, 13–21.

Fernández, A. L., and 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*, 239–46.

Fischer, R., Fontaine, J. R. J., van de Vijver, F. J. R., and van Hemert, D. A. (2009). An examination of acquiescent response styles in cross-cultural research. In A. Gari and K. Mylonas (eds.), *Quod erat demonstrandum: From Herodotus' ethnographic journeys to cross-cultural research* (pp. 137–48). Athens: Pedio Books Publishing.

Fischer, R., and 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., and Schwartz, S. H. (2008). Structural equivalence of the values domain across cultures: Separating sampling fluctuations from meaningful variation. *Journal of Cross-Cultural Psychology*, *39*, 345–65.

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

Harzing, A. (2006). Response styles in cross-national survey research: A 26-country study. *Journal of Cross Cultural Management*, *6*, 243–66.

Ho, D. Y. F. (1996). Filial piety and its psychological consequences. In M. H. Bond (ed.), *Handbook of Chinese psychology* (pp. 155–65). Hong Kong: Oxford University Press.

Hofer, J., Chasiotis, A., Friedlmeier, W., Busch, H., and 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 edn). Thousand Oaks, CA: Sage.

Holland, P. W., and Wainer, H. (eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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

Jahoda, G. (1982). *Psychology and anthropology: A psychological perspective*. London: Academic Press.

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

Johnson, T. P., and van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. R. van de Vijver and P. P h. Mohler (eds.), *Cross-cultural survey methods* (pp. 195–204). New York: Wiley.

Kiers, H. A. L. (1990). *SCA: A program for simultaneous components analysis*. Groningen: 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 and H. Wainer (eds.), *Differential item functioning* (pp. 349–64). Hillsdale, NJ: Erlbaum.

Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N., and 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 and J. Allik (eds.), *The five-factor model across cultures* (pp. 105–26). New York: Kluwer Academic/Plenum Publishers.

McCrae, R. R., and Allik, J. (eds.). (2002). *The Five-Factor Model of personality across cultures*. New York: Kluwer Academic/Plenum Publishers.

Meiring, D., van de Vijver, F. J. R., Rothmann, S., and 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., and Allen, N. J. (1991). A three-component conceptualization of organizational commitment. *Human Resource Management Review, 1*, 61–89.

Patel, V., Abas, M., Broadhead, J., Todd, C., and Reeler, A. (2001). Depression in developing countries: Lessons from Zimbabwe. *British Medical Journal, 322*, 482–84.

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*, 737–56.

Poortinga, Y. H., and Van de Vijver, F. J. R. (1987). Explaining cross-cultural differences: Bias analysis and beyond. *Journal of Cross-Cultural Psychology, 18*, 259–82.

Poortinga, Y. H., van de Vijver, F. J. R., Joe, R. C., and van de Koppel, J. M. H. (1987). Peeling the onion called culture: A synopsis. In Ç. Kağitçibaşi (ed.), *Growth and progress in cross-cultural psychology* (pp. 22–34). Lisse: Swets & Zeitlinger.

Poortinga, Y. H., and van der Flier, H. (1988). The meaning of item bias in ability tests. In S. H. Irvine and J. W. Berry (eds.), *Human abilities in cultural context* (pp. 166–83). Cambridge: Cambridge University Press.

Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., and Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology, 53*, 541–62.

Sandal, G. M., van de Vijver, F. J. R., Bye, H. H., Sam, D. L., Amponsah, B., Cakar, N., Franke, G., Ismail, R. Kai-Chi, C., Kjellsen, K., and Kosic, A. (in preparation). *Intended Self-Presentation Tactics in Job Interviews: A 10-Country Study.*

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., and King, D. W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment, 13*, 442–53.

Sireci, S. (2009). Evaluating test and survey items for bias across languages and cultures. In D. M. Matsumoto and F. J. R. van de Vijver (eds,), *Cross-cultural research methods in psychology.* New York: Cambridge University Press.

Smith, T. (2003). Developing comparable questions in cross-national surveys. In J. A. Harkness, F. J. R. van de Vijver and P. Ph. Mohler (eds.), *Cross-cultural survey methods* (pp. 69–91). New York: Wiley.

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.

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

Tanaka-Matsumi, J., and Draguns, J. G. (1997). Culture and psychotherapy. In J. W. Berry, M. H. Segall and Ç. Kağitçibaşi (eds.), *Handbook of cross-cultural psychology* (vol. III, pp. 449–491). Needham Heights, MA: Allyn and Bacon.

Tashakkori, A., and Teddlie, C. (eds.). (2003). *Handbook on mixed methods in the behavioral and social sciences.* Thousand Oaks, CA: Sage Publications.

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.

(2002). Inductive reasoning in Zambia, Turkey, and The Netherlands: Establishing cross-cultural equivalence. *Intelligence, 30*, 313–351.

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

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

Van de Vijver, F. J. R., and Leung, K. (2009). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. M. Matsumoto and F. J. R. van de Vijver (eds.), *Cross-cultural research methods in psychology*. New York: Cambridge University Press.

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

Van de Vijver, F. J. R., and Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology*, *33*, 141–56.

Van Hemert, D. A., van de Vijver, F. J. R., Poortinga, Y. H., and Georgas, J. (2002). Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. *Personality and Individual Differences*, *33*, 1229–49.

Van Herk, H., Poortinga, Y. H., and 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*, 346–60.

Van Leest, P. F. (1997). Bias and equivalence research in the Netherlands. *European Review of Applied Psychology*, *47*, 319–29.

Van Schilt-Mol, T. M. M. L. (2007). *Differential Item Functioning en itembias in de Cito-Eindtoets Basisonderwijs* [*Differential item functioning and item bias in the Cito Eindtoets Basisonderwijs*]. Amsterdam: Aksant.

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

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*, 525–50.