ONLINE READINGS IN PSYCHOLOGY AND CULTURE
International Association for Cross-Cultural Psychology

Unit 2 *Theoretical and Methodological Issues*
Subunit 2 *Methodological Issues in Psychology and Culture*

Article 8

6-1-2012

Bias and Equivalence in Cross-Cultural Research

Jia He
Tilburg University, j.he2@tilburguniversity.edu

Fons van de Vijver
Tilburg University, fons.vandevijver@tilburguniversity.edu

Recommended Citation

He, J., & van de Vijver, F. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture, 2*(2). http://dx.doi.org/10.9707/2307-0919.1111

Bias and Equivalence in Cross-Cultural Research

Abstract

Bias and equivalence are key concepts in the methodology of cross-cultural studies. Bias is a generic term for any challenge of the comparability of cross-cultural data; bias leads to invalid conclusions. The demonstration of equivalence (lack of bias) is a prerequisite for any cross-cultural comparison. we first describe considerations that are relevant when choosing instruments in a cross-cultural study, notably the question of whether an existing or new instrument is to be preferred.We then describe the definition, manifestation, and sources of three types of bias (construct, method, and item bias), and three levels of equivalence (construct, measurement unit, and full score equivalence). We provide strategies to minimize bias and achieve equivalence that apply either to the design, implementation, or statistical analysis phase of a study. The need to integrate these strategies in cross-cultural studies is emphasized so as to increase the validity of conclusions regarding cross-cultural similarities and differences and rule out alternative explanations of cross-cultural differences.

Introduction

This paper deals with methodological aspects of cross-cultural research, focusing on two key concepts: bias and equivalence. Bias refers to nuisance factors that jeopardize the validity of instruments applied in different cultures. Equivalence refers to the level of comparability of scores across cultures. Some countries use kilometers to measure road distances whereas other countries use miles. Distances in kilometers and miles cannot be directly compared. However, a simple formula (1 mile is about 1.6 km) allows us to convert one scale to the other. After this conversion, the data are comparable (equivalent) and distances can be compared across countries. The example illustrates two important characteristics of bias and equivalence. Firstly, bias does not refer to random errors but to systematic measurement anomalies that are expected to be replicable if a study were to be repeated. Secondly, equivalence is a characteristic of cross-cultural comparisons and not an intrinsic property of instruments; both kilometers and miles are adequate units to measure distances and any lack of equivalence issues arise only in the comparison of both.

Carefully dealing with methodological challenges of cross-cultural research usually involves the minimization of bias and the evaluation of equivalence. Such a combined approach is the foundation to solve challenges such as determining whether an instrument can be used in a different cultural context and whether the comparability of data is ensured in studies concerning multiple cultures. In the remainder of this paper, we first describe considerations that are relevant when choosing instruments in a cross-cultural study, notably the question of whether an existing or new instrument is to be preferred. We then describe and illustrate different types of bias and equivalence. Finally, we provide guidelines to minimize bias and achieve equivalence at different stages of cross-cultural research.

Instrument Choice in Cross-Cultural Studies

An important question to consider in the initial stages of a project involves the choice of instruments. We argue that there are three options (Harkness, Van de Vijver, & Johnson, 2003; Van de Vijver, 2003; Van de Vijver & Leung, 1997).

Adoption

The first option, called *adoption*, amounts to a close translation of an instrument in a target language. This option is the most frequently chosen in empirical research because it is simple to implement, cheap, has a high face validity, and retains the opportunity to compare scores obtained with the instrument across all translations. However, adopting instruments can be a “quick and dirty” solution. The approach has an important limitation, as it can only be used when the items in the source and target language versions have an adequate coverage of the construct measured. So, this option is available if (and only if) the construct and instrument features (e.g., instructions and items) are taken to be adequate in all cultural groups involved.

Adaptation

The second option is labeled *adaptation*. It usually amounts to a combination of a close translation of some stimuli and a change of other stimuli when a close translation would be inadequate for linguistic, cultural, or psychometric reasons. The option has become so popular that adaptation has become the generic term to refer to the translation process of psychological instruments (Hambleton, Merenda, & Spielberger, 2005). The use of the term flags a significant change in the way of thinking about the translation process. Whereas in the past the process of rendering an instrument in another language was mainly viewed as a linguistic task, it has become more common to view this process as requiring more than linguistic skills such as knowledge of the target culture so as to be able to evaluate the psychological relevance of the instrument in the new context.

Assembly

The third option is called *assembly*. It involves the compilation of a new instrument. It is the only choice that remains if adopting or adapting an instrument will not produce an instrument with a satisfactory linguistic, cultural, and psychometric accuracy. An assembly maximizes the cultural suitability of an instrument, but it will preclude any numerical comparisons of scores across cultures.

Selection criteria

Depending on the instrument and target culture, any of the three options (adoption, adaptation, and assembly) may be the best choice. If the aim is to maximize the opportunities for statistical comparisons in a study, adoption is the simplest choice. If the aim is to maximize the ecological validity of the instrument (i.e., to adequately measure the construct in a target culture), an adaptation or assembly is preferable. Statistical tools such as item response theory and structural equation modeling can deal with instruments that are not completely identical across cultures (Van de Vijver & Leung, 1997). However, if the number of culture-specific items is large, the comparability of the construct or of the remaining items may be problematic and opportunities for cross-cultural comparisons are limited. So, maximizing local validity and cross-cultural comparability can be incompatible goals.

Taxonomy of Bias

Bias occurs when score differences on the indicators of a particular construct do not correspond to differences in the underlying trait or ability (Van de Vijver & Tanzer, 2004). This incomplete correspondence means in practice that whereas a response in one culture represents a target construct (e.g., conscientiousness), responses in another country are due to other constructs (e.g., social desirability) or additional constructs (a combination of conscientiousness and social desirability). We argue that there are three types of bias, depending on whether the invalidity comes from the theoretical construct, measurement instrument, or specific items. These types are called construct bias, method bias, and item bias (also called differential item functioning) (Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004).

Construct Bias

Construct bias indicates that the construct measured is not identical across cultures. It can occur when there is only a partial overlap in definition of the construct across cultures, or when not all relevant behaviors associated with the construct are present and properly sampled in each culture (Van de Vijver & Poortinga, 1997). For example, happiness has a different focus in western and non-western countries. According to Uchida, Norasakkunkit, and Kitayama (2004), North Americans tend to derive happiness from personal achievements through maximizing positive affect experiences, whereas East Asians tend to define happiness as interpersonal connectedness with balanced experiences of positive and negative affect. In cases like this, assessing the psychological meaning of happiness requires multiple aspects of happiness to be taken into consideration; the outcome of the assessment should acknowledge the incompleteness of overlap of the construct. Another example is the expression of depression in different cultures. It was found that Chinese depressed outpatients mentioned somatic symptoms as the major complaints, whereas their Australian counterparts stressed depressed mood and cognitive anxiety symptoms more often (Parker, Cheah, & Roy, 2001). This finding highlights the need to employ culture-sensitive measures when administering a depression inventory for patients from various cultures.

Method Bias

Method bias is a generic term for nuisance factors that derive from the sampling, structural features of the instrument, or administration processes.

Sample bias

Sample bias results from incomparability of samples due to cross-cultural variation in sample characteristics that have a bearing on target measures, such as confounding cross-cultural differences in education levels when testing intelligence, variations in urban or rural residency, or in affiliation to religious groups. The ideal situation is to randomly sample culturally representative respondents; yet, due to resources and accessibility restraints, it is rarely accomplished. Many cross-cultural studies use college students, implicitly assuming that they constitute matching samples. However, this assumption may be invalid; for example, college education quality and enrolment rates in developed and developing countries differ significantly, which can introduce selection biases in the sampling process. To minimize sampling bias, Boehnke, Lietz, Schreier, and Wilhelm

(2011) suggested that the sampling of cultures should be guided by research goals (e.g., select heterogeneous cultures if the goal is to establish cross-cultural similarity and homogenous cultures if looking for cultural differences). When participants are recruited using convenience sampling, the generalization of findings to their population can be problematic; the distribution of the target variable is to guide optimal sampling. If the strategy to find matched samples does not work, it may well be possible to control for factors that induce sample bias (such as the measurement of educational quality when assessing intelligence so that a statistical correction for the confounding differences can be achieved).

Instrument bias

Instrument bias involves problems deriving from instrument characteristics, such as stimulus familiarity (in cognitive and educational tests) and response styles (in personality and attitude inventories). Cultures tend to have different levels of familiarity with stimulus materials (e.g., pictures taken in one culture may be not easily identified by members of other cultures), response modes (e.g., differences in familiarity with computers in computer-assisted assessment), or response procedures (e.g., working with multiple choice formats). Such cross-cultural differences in background characteristics tend to influence the scores on target measures. Malda, Van de Vijver, and Temane (2011) confirmed the influence of content familiarity in their study of a cognitive test in two cultures in South Africa. These authors developed test versions with an item content derived from either the Afrikaans (White) or Tswana (Black) culture in South Africa. They found that children from either culture performed better when the version was designed for their own group. Another example was described by Demetriou et al. (2005); they found that Chinese children outperformed Greek children on tasks of visual-spatial processing, which could be attributed to Chinese children's intensive visual-spatial practice involved in learning to write Chinese. To tackle biases arising from stimulus familiarity, tests should be locally adapted (e.g., Malda et al., 2008).

Response styles

Response styles refer to a systematic tendency to use certain categories of the answering scale on some basis other than the target construct (Cronbach, 1950). Acquiescence, the tendency to agree rather than disagree to propositions in general (Lentz, 1938), is one of the most prevalent response styles. Studies have shown that acquiescence is more frequently endorsed by people with low socioeconomic status from collectivistic cultures (Harzing, 2006; Smith & Fischer, 2008). Evidence suggests that the number of Likert points in rating scales may induce different levels of response styles (e.g., Hui & Triandis, 1989; Weijters, Cabooter, & Schillewaert, 2010). For example, Weijters et al. (2010) found that acquiescence increases when adding a midpoint in the response anchors. Different standardization methods using information of means and standard deviations in individuals or cultures have been proposed to control for response styles (see Fischer, 2004, for a review). To determine the extent to which scores are influenced by response styles, correlations can be computed between the corrected scores based on within-individual or within-cultural standardization of the raw scores. Differences in the size of correlations may point to the salience of response styles in the data. The GLOBE leadership project applied a new approach to detect response styles, in which standardized scores were used to predict the raw scores in a regression analysis, and then the raw scores were compared with the predicted scores in *t* tests to identify cultures exhibiting substantial response styles (Hanges, 2004). Other ways of dealing with response styles involve the computation of response style scores, such as computing the proportion of items that are endorsed as a measure of acquiescence or the proportion of extreme responses as a measure of extremity scoring (Van Dijk, Datema, Piggen, Welten, & Van de Vijver, 2009). The computation of social desirability scores is usually more involved?? as it requires the administration of an instrument to measure the construct (whereas acquiescence and extremity can be computed on the basis of measures of other constructs). However, caution is needed in the use of corrections for these response styles; methods to adjust for response styles may remove genuine cross-cultural differences if individual or cross-cultural differences in scores are not just based on response styles but on a combination of response styles and genuine differences.

Administration bias

A final type of method bias is *administration bias*. This type of bias can come from administration conditions (e.g., data collection modes, class size), ambiguous instructions, interaction between administrator and respondents (e.g., halo effects), and communication problems (e.g., language difference, taboo topic). Depending on the constructs of interest, the data collection mode (e.g., paper-and-pencil mode versus online survey) may show differential levels of social desirability. Dwight and Feigelson (2000) found that impression management (one dimension of social desirability) was lower in online assessment. Another case is the interviewer effect; Davis and Silver (2003) revealed that, in answering questions regarding political knowledge, African American respondents got fewer answers right when interviewed by a European American interviewer than by an African American interviewer. These administration conditions that can lead to bias should be taken into consideration before the field work.

In general, method bias tends to have a global influence on cross-cultural score differences (e.g., mean scores of measures vulnerable to social desirability tend to be shifted upwards or downwards). If not appropriately taken into account in the analysis of data, method bias can be misinterpreted as real cross-cultural differences.

Item Bias

An item is biased when it has a different psychological meaning across cultures. More precisely, an item of a scale (e.g., measuring anxiety) is said to be biased if persons with the same trait, but coming from different cultures, are not equally likely to endorse the item (Van de Vijver & Leung, 1997). Item bias can arise from poor translation, inapplicability of item contents in different cultures, or from items that trigger additional traits or have words with ambiguous connotations. For instance, certain words (e.g., the English word "distress") or expressions in one language (e.g., "I feel blue") may not have equivalents in a second language, which challenges the translations of an instrument. When applying the Marlowe-Crowne Social Desirability Scale in different cultures, the item "I never make a long trip without checking the safety of my car" does not apply to most college students in developing countries (Van de Vijver & Meiring, 2011, March). As a result, this item introduces bias and endangers the comparison of scores at item level.

Taxonomy of Equivalence

The taxonomy of bias presented in the previous section dealt with systematic errors in cross-cultural studies. The taxonomy of equivalence, presented below, addresses the implications of bias on the comparability of constructs and scores. More specifically, equivalence is related to the measurement level at which scores obtained in different cultural groups can be compared. Van de Vijver and Leung (1997) proposed a hierarchical classification of equivalence, distinguishing construct equivalence, measurement unit equivalence, and full score equivalence.

Construct Equivalence

There is construct equivalence in a cross-cultural comparison if the same theoretical construct is measured in each culture. Without construct equivalence, there is no basis for any cross-cultural comparison; it amounts to comparing apples and oranges. As argued by Berry (1969), construct equivalence is a prerequisite for cross-cultural comparison. Researchers need to explore the structure of the construct and adequacy of sampled items. When a construct does not have the same meaning across the cultures in a study, researchers have to acknowledge the incompleteness of conceptualization and compare the equivalent sub-facets. For example, filial piety, as a socially approved virtue, contains attributes of respecting, caring for, and loving one's parents in most cultures; however, filial piety in the Chinese culture is broader and also involves obedience and unlimited responsibility to parents, which may amount to taking care of parents when they grow old and needy (Dai & Dimond, 1998). To compare filial piety among western and non-western cultures, researchers should constrain the construct to the sub-facets of filial piety that are recognized in all cultures and acknowledge that, in order to retain comparability, the construct is incompletely covered in one of the cultures.

Measurement Unit Equivalence (Metric Equivalence)

Measurement unit equivalence means that measures of interval or ratio level have the same measurement unit but different origins. With metric equivalence, scores can be compared within cultural groups (e.g., male and female differences can be tested in each group), and mean patterns and correlations across cultural groups, but scores cannot be compared directly across groups. A case in point is the distance being measured by kilometers and miles in the example at the beginning. Distances measured by kilometers can be compared directly, so can distances measured by miles, yet without converting the two measurements to the same origin, a valid cross-group comparison is impossible.

Full Score Equivalence (Scalar Equivalence)

Full score equivalence, the highest level of equivalence, implies that scales have the same measurement unit and origins. In this case, scores obtained are bias free and thus can be compared directly. Analyses of variance and *t* tests to examine cross-cultural differences in means are appropriate for (and only for) this level of equivalence.

It should be noted that in order to achieve construct equivalence, construct bias should be addressed; method and item bias may not influence construct equivalence, but they jeopardize measurement unit and full score equivalence. In the next section, we provide some guidelines to deal with bias in cross-cultural research.

Steps to Reduce Bias and Establish Equivalence

It is becoming more customary to not only report reliability and validity (DeVellis, 2002; Nunnally, 1978), but also to demonstrate equivalence in cross-cultural research. We view this practice as recommendable, because such an analysis can help to bolster conclusions about cross-cultural similarities and differences. Van de Vijver and Tanzer (2004) proposed a detailed scheme to identify and deal with different biases, as shown in Table 1. To tackle biases, we highlight the most important strategies to consider in the following three research stages: design, implementation, and analysis. Minimizing bias in cross-cultural studies usually amounts to a combination of strategies: integrating design, implementation, and analysis procedures. A detailed instruction on cross-cultural survey guidelines can be found at http://ccsg.isr.umich.edu/index.cfm.

At the Design Stage

To ensure construct equivalence in a cross-cultural comparative study, two comparability-driven approaches to design a study have been recommended: decentering and convergence (Van de Vijver & Leung, 1997). Cultural decentering (Werner & Campbell, 1970) means that an instrument is developed simultaneously in several cultures and only the common items are retained for the comparative study; making items suitable for a cross-cultural context in this approach often implies the removal of item specifics, such as references to places and currencies when these concepts are not part of the construct measured. The resulting instrument can be viewed as *adaptation* in the initial stage of item development so that this stage can be followed up by *adoption* when test versions in the target languages are prepared. Large international assessment programs such as PISA (Program of International Student Assessment, details can be found at

Table 1.
Strategies to Reduce Bias in Cross-Cultural Assessment (after Van de Vijver & Tanzer, 2004)

Type of Bias	Strategies
Construct bias	Decentering (i.e., simultaneously developing the same instrument in several cultures) Convergence approach (i.e., independent within-culture development of instruments and subsequent cross-cultural administration of all instruments)
Construct bias and/or method bias	Use of informants with expertise in local culture and language Use samples of bilingual subjects Use of local surveys (e.g., content analyses of free-response questions) Non-standard instrument administration (e.g., think aloud) Cross-cultural comparison of nomological networks (e.g., convergent/discriminate validity studies, monotrait-multimethod studies, connotation of key phrases
Method bias	Extensive training of administrators (e.g., increasing cultural sensitivity) Detailed manual/protocol for administration, scoring, and interpretation Detailed instructions (e.g., with sufficient number of examples and/or exercise) Use of subject and context variables (e.g., educational background) Use of collateral information (e.g., test-taking behavior or test attitudes) Assessment of response styles Use of test-retest, training and/or intervention studies
Item bias	Judgmental methods of item bias detection (e.g., linguistic and psychological analysis) Psychometric methods of item bias detection (e.g., Differential Item Functioning analysis) Error or distracter analysis Documentation of “spare items” in the test manual which are be equally good measures of the construct as actually used test items

http://www.pisa.oecd.org) mostly adopt this approach, in which committee members from participating cultures meet to develop culturally appropriate concepts and measures. In the convergence approach, instruments are developed independently within cultures, and all instruments are subsequently administered in all cultures (Campbell, 1986). Despite the cumbersomeness of the need to administer many instruments, an advantage of the approach is that it captures both universal aspects and cultural specifics of a construct. Instruments developed from the convergence approach are a combination of *assembly* and, subsequently, *adoption* in terms of instrument choice. An example can be found in Cheung, Cheung, Leung, Ward, & Leung (2003). Both the NEO-Five Factor Inventory (NEO-FFI) developed and validated mostly in Western countries and the Chinese Personality Assessment Inventory (CPAI) developed in the Chinese local context were administered to Chinese and Americans. Joint factor analysis of the two personality measures revealed that the Interpersonal Relatedness factor of the CPAI was not covered by the NEO-FFI, whereas the Openness domain of the NEO-FFI was not covered by the CPAI. Consequently, one can expect that merging items from both measures may show a more comprehensive picture of personality.

Before starting the main field work in cross-cultural research, qualitative pilot studies and cognitive interviewing can be used as an informal test of the suitability of instruments and their application procedure. Although these procedures do not ensure the success of the main study, they provide information about feasibility and comparability, expose potential design flaws, and help researchers to refine the assessment process. For instance, Calderón et al. (2006) carried out a qualitative pilot study on the perceptions of ethnic minorities to participate in clinical research. They identified the shared and ethnic-specific barriers and motivators among African American and immigrant Latinos; with this information, they developed targeted community-based strategies to increase minority participation. Cognitive laboratories, often involving a think-aloud strategy, are frequently applied in educational assessment to provide instant feedback of respondents' understanding of test items (Van Someren, Barnard, & Sandberg, 1994).

At the Implementation Stage

In the implementation process, a standard protocol should be developed and abided by all field researchers. The interaction between administrators and respondents should be carefully monitored. Brislin (1986) stressed selecting the right administrator/interviewers, with whom the respondents feel at ease and do not experience cultural barriers. The proper administration process can contribute to the minimization of various response biases that may result from the uncertainties of cross-cultural encounters. Administrators should have intercultural communication competence, so that they can deal with cultural diversity in a professional manner. To facilitate the data collection, administrators need to give clear instructions with sufficient examples. Careful documentation of the field work as well as feedback from respondents could be collected for further analysis. For instance, combining the nonresponse information from the European Social Survey (http://www.europeansocialsurvey.org) and a detailed interviewer questionnaire, Blom, De Leeuw, and Hox (2011) concluded that systematic country differences in nonresponse could partially be attributed to interviewer characteristics such as contacting strategies.

At the Analysis Stage

Many analytic approaches to detect bias and ensure equivalence have been proposed. We restrict the description here to the utilization of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) for different levels of equivalence. We also briefly present differential item functioning analysis (DIF) for item bias detection.

Exploratory factor analysis

EFA is a useful tool to check and compare factor structures, especially when the underlying dimensions of a construct are unclear. Researchers can apply dimensionality-reduction techniques and take the similarity of underlying dimensions as criterion for the similarity of meaning. Comparisons of multiple groups can be conducted either in a pair-wise or a one-to-all (each cultural group versus the pooled solution) fashion. Target rotations are employed to compare the structure across countries and to evaluate factor congruence, often by means of the computation of Tucker's phi coefficient (Van de Vijver & Poortinga, 2002). This coefficient tests to what extent factors are identical across cultures. Values of Tucker's phi above .90 are usually considered to be adequate and values above .95 to be excellent. Tucker's phi can be computed with dedicated software such as an SPSS routine (syntax available from Van de Vijver & Leung, 1997, and www.fonsvandevijver.org).

Confirmatory factor analysis

A more refined and theory-driven way of examining equivalence is through confirmatory factor analysis (CFA, also known as structural equation modeling). If a CFA model shows an acceptable fit, it means that the factor structure assumed cannot be rejected, thus different levels of equivalence may be established (important CFA models and their implications for different levels of equivalence are presented below). More sophisticated than EFA, CFA uses covariance matrix information to test hierarchical models. It can be carried out with software such as AMOS and Mplus (Byrne, 2001, 2010). The model fit is evaluated by Chi-square tests and indices, such as the Tucker Lewis Index (acceptable above .90 and excellent above .95), Root Mean Square Error of Approximation (acceptable below .06 and excellent below .04), and Comparative Fit Index (acceptable above .90 and excellent above .95) (Kline, 2010; for an example see Campos, Zucoloto, Bonafe, Jordani, & Maroco, 9011).

If we want to test whether the same one-factor model holds in each culture, a series of nested models are usually tested for identity (called invariance in confirmatory factor analysis) (e.g., Cheung & Rensvold, 2002). We illustrate five models which give important indications of equivalence; their operationalization and interpretation are presented in Table 2. The *configural invariance model* is a starting point. In this model, the same latent construct with the same indicators are assumed. It is a base for testing the nested models illustrated below. In the *measurement weights model*, factor loadings on the latent variable are constrained to be equal across cultures. If the multigroup confirmatory factor analysis yields a satisfactory fit, the construct under investigation can be said to have construct equivalence and that the construct has the same connotation across groups. In the *intercept invariance model*, items are constrained to have the same intercept (latent mean) across cultures. The working assumption is that individuals who have the same score on the latent construct would obtain the same score on the observed variable regardless of cultural membership. It is used to detect item bias; if this model shows a satisfactory fit, it

Table 2.
Nested Models in Multigroup Confirmatory Factor Analysis

Hierarchical Models	Operationalization	Interpretation of level of equivalence
1. Configural invariance	Same pattern of observed and latent constructs	Same latent constructs are measured, using the same indicators (no metric equivalence)
2. Measurement weights	Factor loadings in the measurement part in each cultural group are identical	Same latent factor(s) is/are measured across groups, indicating construct and metric equivalence
3. Intercept invariance	Items have the same intercept (latent mean) across cultures	All items represent the same between-group difference, indicating free of item bias and full score equivalence
4. Structural residual	The error variance of the latent factor is identical	The range of scores on the latent factor does not vary across cultures, indicating full score equivalence
5. Measurement residuals	Error variances of the observed items are identical	Groups use the same range of the construct continuum, indicating full score equivalence

can be assumed that there is no item bias. A poor fit alerts researchers to check anomalous items that relate to the latent scores in different manners. The acceptance of the *structural residual model*, in which the error of the latent variable is fixed equal across cultures, indicates that measurement unit equivalence is guaranteed. The *measurement residuals model*, the most restricted model, specifies the same error variance for each and every item. A satisfactory fit of this model represents full score equivalence and it lays a solid foundation for cross-cultural comparison.

Item bias or differential item functioning analysis

When all possible precautions for bias presented in Table 1 are taken but factor analysis still suggests lack of equivalence among cultural groups, it may be useful to investigate to what extent anomalous items could be responsible. Differential item functioning (DIF) analysis can identify such anomalous items. DIF indicates that respondents from different cultures show differing probabilities of correctly solving or endorsing the item after matching on the underlying ability that the item is intended to measure (Zumbo, 1999). In this analysis, scales should be unidimensional (for multidimensional constructs, DIF analyses can be performed per dimension).

As mentioned in the preceding section, a poor fit of the intercept invariance model in CFA suggests that items are biased. A simple estimation of DIF, based on analysis of variance, can be done in three steps. First the total scores of a unidimensional scale, irrespective of cultures, are computed. Second, the total scores are divided into several levels based on the range. Third, an ANOVA is performed, in which culture and score level serve as the independent variables and item scores are the dependent variable. A significant effect of culture and the interaction between culture and score level points to item bias; such a finding implies that scores on that item cannot be directly compared across cultures. A closer inspection of the items may then reveal whether there is a translation issue, whether the item is unrelated to the underlying trait in one culture, or whether the item identifies an interesting cross-cultural difference that requires further scrutiny. For example, Kalaycioglu and Berberoglu (2011) examined gender bias in university entrance exams in Turkey and they found that numerical and symbolic representations used in item content were the two sources of DIF favoring male students, whereas routine algorithmic calculations could produce DIF against males.

More advanced DIF procedures can be found in item response theory, logistic regression, and Mantel-Haenszel tests (Osterlind & Everson, 2009), which are beyond the scope of the current paper. A handbook of DIF with SPSS syntax and examples is available at http://www.educ.ubc.ca/faculty/zumbo/DIF (see also Zumbo, 1999).

Conclusion

We have described the choices of instruments, various forms of bias and equivalence, and ways of addressing issues of bias in this paper. All sorts of bias can have hazardous effects on cross-cultural comparisons. We hope this article will help readers to recognize the importance of bias and equivalence issues in cross-cultural research, utilize the strategies outlined, and refrain from making statements about cross-cultural similarities and differences when proper methodological prescriptions have been not been heeded.

References

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4,* 119-128. doi:10.1080/00207596908247261

Boehnke, K., Lietz, P., Schreier, M., & Wilhelm, A. (2011). Sampling: The selection of cases for culturally comparative psychological research. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 101-129). New York: Cambridge University Press.

Blom, A. G., de Leeuw, E. D., & Hox, J. J. (2011). Interviewer effects on nonresponse in the European Social Survey. *Journal of Official Statistics, 27,* 359-377.

Brislin, R. W. (1986).The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research (*Vol. 8, pp. 137-164). Thousand Oaks, CA: Sage Publications.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Byrne*,* B. M. *(2010). Structural equation modeling* with *Mplus*: *Basic concepts*, *applications*, and *programming*. New York: Routledge.

Calderón, J. L., Baker, R. S., Fabrega, H., Conde, J. G., Hays, R. D., Fleming, E., & Norris, K. (2006). An ethno-medical perspective on research participation: A qualitative pilot study. *Medscape General Medicine, 8,* 23.

Campbell, D. T. (1986). Science's social system of validity-enhancing collective belief change and the problems of the social sciences. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science: Pluralities and subjectivities* (pp. 108-135). Chicago, IL: University of Chicago Press.

Campos, J., Zucoloto, M. L., Bonafe, F. S. S., Jordani, P. C., & Maroco, J. (2011). Reliability and validity of self-reported burnout in college students: A cross randomized comparison of paper-and-pencil vs. online administration. *Computers in Human Behavior, 27*, 1875-1883. doi:10.1016/j.chb.2011.04.011

Cheung, F. M., Cheung, S. F., Leung, K., Ward, C., & Leong, F. (2003). The English version of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology, 34,* 433-452. doi:10.1177/0022022103034004004

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. doi:10.1207/s15328007sem0902_5

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10,* 3-31.

Dai, Y. T., & Dimond, M. F. (1998). Filial piety: A cross-cultural comparison and its implications for the well-being of older parents. *Journal of Gerontological Nursing, 24,* 8-13.

Davis*,* D. W., & Silver*,* B. D. *(*2003*).* Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science, 47,* 33-45. doi:10.1111/1540-5907.00003

Demetriou, A., Kui, Z. X., Spanoudis, G., Christou, C., Kyriakides, L., & Platsidou, M. (2005).The architecture, dynamics, and development of mental processing: Greek, Chinese, or universal? *Intelligence, 33,* 109-141. doi:10.1016/j.intell.2004.10.003

DeVellis, R. F. (2002). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Dwight, S. A., & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement, 60*, 340-360. doi:10.1177/00131640021970583

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

Hambleton, R. K., Merenda P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational tests and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

Hanges, P. (2004). Response bias correction procedure used in GLOBE. In R. J. House, P. I. Hanges, M. Javidan, P. J. Dorfman, & Gupta (Eds.), *Culture, leadership, and organizations: The GLOBE study of 62 culture* (pp. 737-752). Thousand Oaks, CA: Sage Publications.

Harkness, J. A., Van de Vijver, F. J. R., & Johnson, T. P. (2003). Questionnaire design in comparative research. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 19-34). New York: Wiley.

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management, 6*, 243-266. doi:10.1177/1470595806066332

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296-309. doi:10.1177/0022022189203004

Kalaycioglu, D. B., & Berberoglu, G. (2011). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment, 29*, 467-478. doi:10.1177/ 0734282910391623

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Lentz, T. F. (1938). Acquiescence as a factor in the measurement of personality. *Psychological Bulletin 35*, 659.

Malda, M., Van de Vijver, F. J. R., Srinivasan, K., Transler, C., Sukumar, P., & Rao, K. (2008). Adapting a cognitive test for a different culture: An illustration of qualitative procedures. *Psychology Science Quarterly, 50*, 451-468.

Malda, M., Van de Vijver, F. J. R., & Temane, M. Q. (2011). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence, 38*, 582-595. doi:10.1016/j.intell.2010.07.004

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Osterlind, S. J., & Everson, H. T. (Eds.). (2009). *Differential item functioning.* Thousand Oaks, CA: Sage.

Parker, G., Cheah, Y. C., & Roy, K. (2001). Do the Chinese somatize depression? A cross-cultural study. *Social Psychiatry and Psychiatric Epidemiology, 36*, 287-293. doi:10.1007/s001270170046

Smith, P. B., & Fischer, R. (2008). Acquiescence, extreme response bias and culture: A multilevel analysis. In F. J. R. Van de Vijver, D. A. van Hemert & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 285-314). New York: Taylor & Francis Group/Lawrence Erlbaum Associates.

Uchida, Y., Norasakkunkit, V., & Kitayama, S. (2004). Cultural constructions of happiness: Theory and empirical evidence. *Journal of Happiness Studies, 5*, 223-239. doi:10.1007/s10902-004-8785-9

Van de Vijver, F. J. R. (2003). Test adaptation/translation methods. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 960-964). Thousand Oaks, CA: Sage.

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., & Meiring, D. (2011, March). *Social desirability among Blacks and Whites in South Africa*. Paper presented at Cross-Cultural Psychology Symposium at Tilburg University, the Netherlands.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29-37. doi:10.1027/1015-5759.13.1.29

Van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology, 33,* 141-156. doi: 10.1177/ 0022022102033002002

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 54,*119-135. doi:10.1016/j.erap.2003.12.004

Van Dijk, T. K., Datema, F., Piggen, A.-L. J. H. F., Welten, S. C. M., & Van de Vijver, F. J. R. (2009). Acquiescence and extremity in cross-national surveys: Domain dependence and country-level correlates. In A. Gari & K. Mylonas (Eds.), *Quod erat demonstrandum: From Herodotus' ethnographic journeys to cross-cultural research* (pp. 149-158). Athens, Greece: Pedio Books Publishing.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think-aloud method: A practical guide to modeling cognitive processes*. San Diego, CA: Academic Press Ltd.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*, 236-247. doi:10.1016/ j.ijresmar.2010.02.004

Werner, O., & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of cultural anthropology* (pp. 398-419). New York: American Museum of National History.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-Type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

About the Authors

Jia He is a Ph.D. researcher at the Department of Cross-Cultural Psychology in Tilburg University, the Netherlands. She obtained her M.A. degree in Intercultural Communication from Shanghai International Studies University, China. Her current research includes the psychological meaning of survey response styles across cultures, values, social desirability, and other methodological aspects of cross-cultural studies. She is also keen in research methods such as structural equation modeling and multilevel analysis.

Webpage: http://www.tilburguniversity.edu/webwijs/show/?uid=j.he2
E-mail address: j.he2@tilburguniversity.edu

Fons van de Vijver is professor of cultural psychology at Tilburg University, the Netherlands, and Professor Extraordinary at North-West University, South Africa and the University of Queensland, Australia. He obtained a PhD from Tilburg University in 1991. The study dealt with cross-cultural differences and similarities in inductive reasoning in Zambia, Turkey, and the Netherlands. He has written over 350 publications, mainly on cognition, acculturation, multiculturalism, and methodological aspects of cross-cultural studies (how can we design and analyze cross-cultural studies so as to maximize their validity?). With Kwok Leung from Hong Kong, he wrote a book on cross-cultural research methods (1997, Sage). He is the former Editor of the Journal of Cross-Cultural Psychology. He is the current President of the European Association of Psychological Assessment.

Webpage: http://www.fonsvandevijver.org,
http://www.tilburguniversity.nl/webwijs/show/?uid=fons.vandevijver
and http://www.psy.uq.edu.au/directory/index.html?id=1955.
E-mail address: fons.vandevijver@tilburguniversity.edu and fons.vandevijver@uq.edu.au

Discussion Questions

1. Select a study in which instruments were applied in different countries from a recent issue of the *Journal of Cross-Cultural Psychology* and determine how the author(s) controlled for bias.
2. Suppose that you conduct a cross-cultural study in which conformity is measured in the US, South Africa, and Japan. How can sources of method bias be controlled in cross-cultural studies in this study? Discuss procedures at both the design and analysis stage.
3. A researcher is interested in comparing levels of depression across several countries, using translations of the Deck Depression Inventory (which is a widely used inventory to assess depression) in the various countries. What are the main issues in terms of bias and equivalence of such a study?
4. A researcher is interested in comparing levels of depression across several countries, using an indigenous approach in which interviews are held with individuals from the target cultures which are then used to formulate items. The items are partly identical across the countries and partly different. What are the main issues in terms of bias and equivalence of such a study?
5. What kind of equivalence would be most important for a study that tries to establish whether extroversion has the same meaning in Morocco, Japan, and the Philippines?

6. Suppose that you want to compare two countries on individualism—collectivism and that the samples in one country has on average a higher level of education than the sample in the other country. Discuss how this difference could challenge your findings and how you could try to disentangle educational and cultural differences.

7. Discuss strengths and weaknesses of judgmental procedures to evaluate differential item functioning.

8. Do you expect social desirability to be higher in Sweden or in China? Motivate your answer.