Towards an integrated analysis of bias in cross-cultural assessment

van de Vijver, F.J.R.; Poortinga, Y.H.

Published in:
European Journal of Psychological Assessment

Publication date:
1997

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
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*(1), 29-37.

European Journal of Psychological Assessment, Vol. 13, Issue 1, pp. 29–37

Towards an Integrated Analysis of Bias in Cross-Cultural Assessment

Fons J. R. van de Vijver[1] and Ype H. Poortinga[1,2]

[1]Tilburg University, The Netherlands, [2]University of Leuven, Belgium

A central methodological aspect of cross-cultural assessment is the interpretability of intergroup differences: Do scores obtained by subjects from different cultural groups have the same psychological meaning? Equivalence (or the absence of bias) is required in making valid cross-cultural comparisons. As cross-cultural comparisons are becoming increasingly popular and important, the problem of bias and its detection is receiving increased attention from researchers. Three kinds of bias are discussed and illustrated, namely construct bias, method bias, and item bias (or differential item functioning). Methods to identify bias are reviewed. An overview is given of common sources of each kind of bias. It is argued that an integrated treatment of all forms of bias is needed to enhance the validity of cross-cultural comparisons. The predominant focus on item bias techniques has the unfortunate implication that construct and method bias are examined insufficiently.

Keywords: Cross-cultural testing, Cross-cultural assessment, construct bias, item bias, detection of bias

Topics in Cross-Cultural Assessment

We live in an era of cross-cultural encounters. Previously closed borders have opened up, more commercial companies than ever before operate at an international level, and migration has transformed essentially monocultural societies into multicultural societies. Not surprisingly, the interest in cross-cultural psychology is steadily growing during the last decades. As much as 1% of all publications covered by *Psychological Abstracts* explicitly deals with cross-cultural comparisons nowadays and this number is steadily growing (Van de Vijver & Lonner, in press). These cross-cultural encounters cannot fail to have an impact on assessment. New instruments for educational or job selection need to be valid in a multicultural context and test developers will have to show the adequacy of their instruments. Furthermore, the ability to operate in a multicultural environment will become increasingly important for employees of multinational companies and institutions. The assessment of abilities to work in a multi-cultural context including, among other things, communication skills, flexibility, and cultural sensitivity, may well become standard practice in selection procedures for international management trainees. Skills in cross-cultural assessment will also be required from growing numbers of professional psychologists.

Methodological considerations are important in cross-cultural comparisons. When a psychological instrument developed in one society is applied in a different cultural context, invariance of psychometric properties (reliability and validity) cannot be merely assumed, but has to be empirically demonstrated. Even excellent properties in each separate society are not a sufficient condition for valid cross-cultural comparison. A test of spatial ability may be valid in two cultural populations, but when the test content or the response procedure is more familiar to members of one group, it is likely that intergroup comparisons are invalid. Valid comparison of scores presupposes that these scores have an equal psychological meaning, not only within but also across cultures. When this is the case the score variable is called "free from bias" or "equivalent." Bias is used as a generic term to indicate a lack of correspondence between the observed scores of subjects from different cultural populations and the domain of generalization; for many tests the domain of generalization is the trait or ability that the test is taken to measure (e.g., the subjects' spatial ability) (Poortinga & Malpass, 1986; Van de Vijver & Leung, in press).

Bias can occur for a variety of reasons. The effects can be limited to one or a few items in an instrument, but it is also possible that all items are affected. The analysis of bias is often limited to the detection of item bias. In this paper, the case is made for an integrated analysis in which various forms of bias are scrutinized. Moreover, bias is not seen as an inherent property of the instrument, but as a function of the interpretation of the test scores.

In the first section of this paper three kinds of bias will be considered: construct bias, method bias, and item bias. This section is followed by a brief discussion of methods to identify each form of bias. In the last three sections an integrated approach to the analysis of bias is introduced and illustrated.

Three Forms of Bias

Bias can occur at three levels. First, the construct that is studied can differ to a substantial degree across cultural groups. This is called *construct bias*. An example can be found in intelligence testing. For example, everyday conceptions of intelligence are much broader than the range of topics covered in intelligence tests. Social intelligence, which includes characteristics such as knowing one's role in the family, and the ability to deal with socially complicated situations, forms part of everyday conceptions, but is usually not covered in psychological tests (Serpell, 1993; Sternberg, 1985; Super, 1981). In this line of reasoning, intelligence tests are better seen as measurements of scholastic or academic intelligence.

Historically speaking, the construct bias in intelligence tests is well understandable, because Binet was asked to design a psychological instrument for a specific purpose, namely the detection of pupils with learning difficulties at school. As an example in the area of personality, the Chinese concept of "filial piety" can be mentioned; it refers to taking care of one's parents, conforming to their requests, and treating them well. The Chinese concept is much broader than the Western concept of "being a good son or daughter" (Ho, in press). In general, construct bias is likely to appear when test authors from various societies use definitions of the concept under study that do not fully overlap.

Method bias occurs when a cultural factor that is not relevant to the construct studied affects most or all items of a test in a differential way across the cultures studied. Response styles, such as the tendency to use or avoid score extremes on a response scale, are an example (e.g., Hui & Triandis, 1989; Poortinga & Foden, 1975). Differential social desirability can also induce method bias; for example, cross-cultural differences in self-disclosure can seriously threaten score comparisons. Other examples can be found in mental testing. Jensen has argued that the Raven tests are culture-reduced, meaning that the performance on these tests is influenced only to a limited extent by familiarity with the stimulus figures and response format. It is probably accurate to state that figures such as used in the Raven tests are less culturally specific than, for example, the items from the information subtest in the WAIS. However, even simple geometric figures are culturally entrenched and not equally familiar across cultural groups. Exposure of children to such figures will substantially differ across cultures; differential exposure will influence the test results and challenge score comparability more in cross-cultural than in intracultural research.

In addition to stimulus content, response procedures can also induce method bias. Serpell (1979) administered a pattern-copying task to children in the United Kingdom and Zambia. The children's copying skills were assessed using two response media: pencil-drawing and iron-wire modelling, a pastime that is popular among Zambian boys. The British children scored higher than the Zambian children on the drawing task while the Zambian children scored higher on the wire modelling task.

The last form of bias is *item bias* or *differential item functioning* (Berk, 1982; Holland & Wainer, 1993). Because in our framework item bias is one of three sources of bias, the prevailing term "differential item functioning" or *DIF* will not be used here. Item bias refers to anomalies of the instrument that are specific to individual items. An item is biased if persons from different groups with the same score on the construct, commonly operationalized as the score on the instrument, do not have the same expected score on the item (Shepard, Camilli, & Averill, 1981). For an unbiased item, knowledge of the total test score of a person does not contain information on group membership, while for a biased item it does.

Two kinds of item bias have been distinguished (Mellenbergh, 1982). The first one, called uniform bias, occurs when the average item score for examinees with a certain test score is lower in one of the populations across the entire range of test scores. Suppose that a test of global geography contains an item asking for the name of the capital of Japan. A Japanese school pupil is more likely to know the answer than a Greek child independent of the overall performance on the test.

Bias can also be nonuniform. In this case the cross-cultural differences in item difficulty or endorsement rate is not the same across the ability or attitude range. Nonuniform bias points to differential discriminatory power of an item across cultural groups. As an example, suppose that in an international comparison of mathematics achievement, the following item is included:

What is the square root of 25?
a) 4 *b) 5 c) 6 d) none of these

Furthermore, suppose that the concept of square roots has been treated in the curriculum of only one of the countries involved. In this country the item may be appropriate whereas in the other countries the item will hardly discriminate. In the first culture the empirical item characteristic curve will be steep while the curve for B may not increase at all across the score range.

The Detection of Bias

Psychometric procedures form the core of *item bias* analysis. A host of techniques have been developed (Berk, 1982; Hambleton, Clauser, Mazor, & Jones, 1993; Holland & Wainer, 1993; Van de Vijver, 1994). They all have in common that expectations based on the other items are used to assess possible bias in an item. A typical example is the Mantel-Haenszel procedure, proposed by Holland and Thayer (1988), nowadays the most popular technique. The Mantel-Haenszel statistic tests per item whether the proportion of persons with a correct response is the same in each cultural group across various score levels. In the first step of the analysis, the examinees are split according to score groups; the first group will consist of all subjects who correctly solved one item (subjects with a score of zero or a perfect score are excluded from the analyses); in the second group the subjects are placed with a sum score of two, and so on. The maximum number of score groups that can be distinguished will be equal to the number of items minus one. Particularly in groups with extreme scores, the available number of subjects can easily become too small to warrant analysis. In such a case a smaller number of score groups may be used such as a low, medium, and high scoring group. In each of these score groups a 2×2 matrix can be composed per item indicating the number of subjects passing and failing that particular item.

For unbiased items the Mantel-Haenszel statistic follows asymptotically a chi-square distribution with one degree of freedom. Items with significant values are taken to be biased. The Mantel-Haenszel procedure performs quite well in Monte Carlo studies if the sample sizes are at least 200 in each group. However, the statistic is better in detecting uniform than nonuniform bias. If there is reason to assume that several items may show nonuniform bias, the utilization of item response theory could be considered which has separate parameters for item difficulty and item discrimination (Hambleton, Swaminathan, & Rogers, 1991). Uniform bias is found when there are differences between groups on the difficulty parameter while differences in the item discrimination parameter point to nonuniform bias.

If the responses to be analyzed are continuous interval-level scores, an analysis of variance can be applied. The procedure is similar to that for the Mantel-Haenszel statistic. First the sample is split up in score groups. Per item, an analysis of variance is carried out with the item score as dependent variable and culture and score group as independent variables. A significant main effect for culture points to the presence of uniform bias while a significant interaction between culture and score group is taken as evidence of nonuniform bias.

Interaction components in an analysis of variance are infamous for their instability. Research on aptitude treatment interactions in educational psychology (Cronbach & Snow, 1977) and on person situation interactions in personality (Endler & Magnusson, 1976) have a demonstrated lack of replicability. Item bias research is troubled by the same problem (Van de Vijver, 1994).

The assessment of *method bias* requires the collection of additional data beyond the instrument that is being investigated for bias. An obvious way to study the impact of the measurement procedures is their systematic variation across cultural groups. A monotrait-multimethod approach will serve this purpose. Both stimulus content and response procedures can be varied. An observation of cross-cultural differences that are dissimilar across measurement procedures can be seen as evidence for method bias.

One of the most important sources of method bias in mental tests is differential test-wiseness, or stimulus familiarity. This can be elegantly studied by application of monotrait-multimethod procedures and by the repeated administration of the instrument. Many mental tests show higher scores on the second occasion (Wing, 1980). Differential gain patterns point to method bias and retrospectively cast doubt on the validity of the first test administration. Individuals with little test experience can be expected to gain more from repeated test administration. Increases that are larger in non-Western groups than in Western groups have been reported (Kendall, Verster, & Von Mollendorf, 1988; Van de Vijver, Daal, & Van Zonneveld, 1986). Moreover, Ombrédane, Robaye, and Plumail (1956) have studied the predictive validity of the *Raven Progressive Matrices* in a group of unschooled miners. These authors found an increase of the predictive validity of the instrument with repeated test administration. The study neatly demonstrates that method bias can also threaten the validity of an instrument.

The most frequently used techniques for the analysis of method bias are factor analysis and, in recent years, the analysis of covariance structures, especially with the LISREL (Jöreskog & Sörbom, 1993) and EQS (Bentler, 1992) programs.

The need for additional data beyond the instrument for which bias has to be assessed (also mentioned in the discussion of method bias) holds even stronger for the assessment of *construct bias*. When only test data on the instrument under study are available, it is usually impossible to decide about the presence or absence of construct bias. Analysis of construct bias will usually start with a survey of definitions of the concept in the cultural populations under study. In the case of filial piety, this could amount to a listing of all behaviors that persons of a particular cultural group associate with being a good son or daughter. An incomplete correspondence of these behaviors points to construct bias. A test user may also conclude that there is bias in the operational definition of a concept when there is evidence of strong bias effects at item and instrument level that can not be eliminated by removal or reformulation of a few items or a change in the administration procedure. Werner and Campbell (1970) used the term "cultural decentering" for the process of eliminating and changing biased items. If much cultural decentering is needed, the conclusion that there is construct bias becomes unavoidable.

Towards a Balanced Treatment of Bias

The present approach to bias distinguishes anomalies at construct, instrument (method), and item level. A balanced treatment of bias requires an open eye for all these forms of bias. In our view, past efforts to detect and correct bias have often been guided by a one-sided focus on the analysis of bias and possibilities to eliminate the effects. Particularly the item bias tradition suffers from an oversimplified view. In case of a large cultural distance between the cultural groups studied, it is unrealistic to assume that item bias techniques can statistically eliminate all irrelevant differences between the groups. When the *Cattell Culture Fair Intelligence Test* is administered to Bushmen, Kalahari desert dwellers, and to Western examinees, the differences in cultural background and relevant prior knowledge are so massive that item bias techniques cannot adequately be applied. Equal test scores just do not have the same psychological meaning across these groups. When the cultural distance between two groups is less, the size of bias effects will be smaller, but at how small a cultural distance do the effects become negligible?

A second problem with many existing approaches to bias has to do with the inferences that are drawn from scores. Items (or instruments) are not inherently biased or unbiased, but they can be biased given particular inferences that are derived from the scores. Bias is more likely when there are large cultural distances to be bridged. This is also the case when inferences based on test scores refer to more encompassing domains of behavior or broader traits. Going back to an earlier example, the higher scores of the Japanese children on the item asking for the capital of this country validly reflects differential knowledge. However, with such an item interpretation of test score differences as indices of general geographical knowledge, the interpretation is likely to be misleading. This holds even more, if an item like this is included in a test of general intelligence. The same arguments apply for method bias.

Impact or Bias?

Inspired by the item bias tradition, a distinction has been proposed between bias and impact. Bias refers to intergroup differences that are due to measurement artifacts whereas impact refers to valid intergroup differences. Using this terminology, bias analyses can be conceived of as the separation of bias and impact; by eliminating measurement artifacts (i. e., bias) a researcher will gain insight into the valid cross-cultural differences (i. e., impact). A refinement of the instrument is taken to bring us to the real differences. An instrument from which all biased items are removed will allow cross-cultural comparisons.

The distinction is a consequence of the too narrow view of bias noted in the previous section. The weaknesses become clear when the implications of method bias for observed performance differences across groups are considered. Both item and method bias lead to systematic score differences between cultural groups. It is highly unlikely that item (or method) bias will be distributed in such a way that the difference between two groups in the mean test score is not affected; particularly item bias effects will differ from item to item, but almost without exception the effects of bias will systematically favor the cultural group from where the instrument originates. As a consequence, valid intergroup differences in average scores can be confounded with bias. Especially when the cultural distances between the groups are large, it is unrealistic to assume that intergroup differences after the removal of all identifiable item bias are to be accounted for entirely by impact. When a substantial proportion of items show evidence of uniform bias in one and the same direction it is more prudent to question the suitability of the instrument and to check for method and construct bias than to remove the biased items.

When there is reason to suspect the presence of method bias, the application of item bias techniques can lead to paradoxical results. Suppose, in a test administered in two cultural groups method bias is present due to differences in stimulus familiarity. Furthermore, suppose that the stimulus familiarity induces an overall difference of 0.5 standard deviation between the scores of both groups. The familiarity effects can be seen as either having the same value for all items, or as random drawings from some distribution. If the items are selected from a narrow domain and the cultural distance is large, the familiarity parameter can be seen as a constant. In such a data set the bias and impact effects are inextricably confounded and item bias techniques will not retrieve the bias, because it is discernible only at test score level.

Stimulus familiarity may not affect all item scores to the same extent and effects on the various items form some kind of distribution. Suppose, they form a normal distribution with a mean of 0.5 (the overall influence of differential familiarity on the mean score difference) and a standard deviation of 0.5. An item bias analysis will consider the average differential familiarity effect as impact and will flag the items with extreme familiarity scores as biased. There will be two kinds of biased items: those with high values on the differential familiarity parameter and those with low (even negative) values. The removal of the former will enhance the validity of intergroup comparisons. However, removal of the items with a low score on the differential familiarity parameter provide the best estimate of the real group difference, and this will decrease the validity of comparisons.

This paradox arises from the fact that bias is present in the total test score, which serves as the standard in terms of which bias in each separate item is examined. The use of a criterion score that itself is not free from bias, leads to undesirable results.

With construct bias there are similar problems. In this case a valid comparison of scores cannot be achieved at all. Elimination of items that do not fit the overall difference in scores does not improve the validity, because the meaning of the difference is unclear; item bias analyses cannot correct a lack of correspondence in concepts across cultures. In sum, in order to make a valid distinction between bias and impact, item bias analyses are useful though they should be complemented by examinations of construct and method bias.

What Form of Bias Can Be Expected?

Several authors have listed possible problems in cross-cultural assessment. Thus, Hambleton (1994) has generated a list of difficulties in test translations (see also Brislin, 1980); Irvine and Carroll (1980) have developed a set of guidelines for test development; Van de Vijver and Poortinga (1991) have generated an overview of issues in test administration. From a methodological point of view, assessment problems are likely to give rise to bias. What kind of bias can be expected when items are poorly translated? What kind of bias can be expected when the operationalizations are specific to one culture? A brief overview of possible problems in cross-cultural assessment and the bias to which they will lead is presented in Table 1. The overview can only be selective given the multitude of problems that can play a role in cross-cultural assessment but an attempt has been made to list the major problem areas.

The most important reason for the occurrence of construct bias is the unfounded assumption of universality of psychological constructs as conceptualized in a particular theory or tradition. Everyday conceptualizations of psychological constructs can vary across cultural groups. The previously mentioned example of social intelligence illustrates the considerable differences in inclusiveness of concepts that can be found. Another problem leading to construct bias is the incomplete coverage of the psychological construct. By considering only a small sample of all possible, relevant behaviors, one can

Table 1. Overview of kinds of bias and their possible causes.

Kind of bias	Source
Construct	– incomplete overlap of definitions of the construct across cultures – differential appropriateness of item content (e.g., skills do not belong to the repertoire of either cultural group) – poor sampling of all relevant behaviors (e.g., short instruments covering broad constructs) – incomplete coverage of the psychological construct
Method	– differential social desirability – differential response styles such as extremity scoring and acquiescence – differential stimulus familiarity – lack of comparability of samples (e.g., differences in educational background, age, or gender composition) – differences in physical testing conditions – differential familiarity with response procedures – tester effects – communication problems between subject and tester in either cultural group
Item	– poor item translation – inadequate item formulation (e.g., complex wording) – one or a few items may invoke additional traits or abilities – incidental differences in appropriateness of the item content (e.g., topic of item of educational test not in curriculum in one cultural group)

quickly load the dice against a cultural group. If a small set of items is used to measure broad constructs, item particulars will have a relatively large influence on observed cross-cultural differences. A measure of anxiety in which items dealing with physical threat are relatively frequent as compared to items about interpersonal anxiety may show a different pattern of cross-cultural differences than a test in which interpersonal items are relatively frequent and physical threat is hardly represented.

Differential appropriateness of item content can also cause construct bias. Suppose that a coping questionnaire has a subscale to measure avoidance with items such as "When I have a serious problem, I go to the movies to forget the problem." Available and preferred distracting activities differ across cultures. Items about going to the movies, watching television or videos, and listening to music may be adequate in Western studies but will be inadequate in groups where these activities are less common.

Method bias can be brought about by social desirability. Its impact will challenge the validity of intergroup comparisons when its influence is not identical across cultures. Norms about appropriate conduct differ across cultural groups and the social desirability expressed in assessment will vary accordingly. Response styles such as acquiescence and extremity scoring can also differ across cultures; hence, they can also jeopardize the validity of intergroup comparisons.

An important source of method bias in cognitive testing is stimulus familiarity. When educational differences between the groups are large, it will be virtually impossible to find stimulus material that is equally familiar to all individuals. This negative view on the possibility of obtaining adequate stimulus material should not be taken to mean that it is futile to try to develop such stimuli. Cross-cultural differences observed may depend on the nature and entrenchment of the stimulus material. A systematic variation of the stimuli in this respect could provide insight into their cultural loadings.

Another important source of method bias are communication problems between tester and subject. It is not uncommon that testing in multilingual societies takes place in the second, or even third, language of the subjects and testers. Lower skills in these languages can affect test performance. The speededness of a test can also give rise to method bias. Individuals who are not experienced with speeded tests often first have to learn how to find a balance between speed and accuracy. Such individuals often choose for extreme speed thereby neglecting accuracy, or for extreme accuracy neglecting speed of responding. In most speed tests a combination of the two strategies will lead to a better performance. Such a strategy will first have to be mastered.

In field research it can be difficult to achieve the same physical testing conditions in each cultural group. Thus, overcrowded school classes or lack of room for individual testing can threaten similarity of testing conditions. Computerized tests require similar conditions in terms of ambient light and absence of glare; such conditions may be hard to obtain in field settings.

Item bias is a consequence of fairly specific anomalies such as poor translation, lower stimulus or response familiarity, and inapplicability of a specific item in a particular cultural group. Complexity of item wording can also induce item bias, in particular when there are educational differences between the groups. When the wording of an item is complex, it may measure cognitive abilities in addition to the intended psychological construct.

An Example: Test Adaptations

An important domain of application of bias studies are test adaptations (or translations), referring to the use of an instrument in multiple languages and cultures (Hambleton, 1994). It is a common practice to translate an instrument from a source into a target language and to back translate the version of the target language into the source language (Brislin, 1980). The original and back translated versions are compared as a check of the accuracy of the translation. This check is important in test adaptations, though certainly not sufficient to ensure psychological equivalence of the instrument in all groups. All forms of bias mentioned can threaten test adaptations and a sensitivity to all these aspects is required. The effects of cultural differences that are not relevant to the construct studied should be minimized so that scores are determined as much as possible by the construct under study. An extensive description of all issues in test adaptation is beyond the scope of this paper, but a selective sample may highlight the issues involved.

In the first stage of the adaptation, construct bias is of major concern: Is the construct identical for the language groups? The use of local informants, colleagues coming from or with a thorough knowledge of each target culture, or a pilot study in which the meaning of the concept in the target groups is examined are useful means to scrutinize the cross-cultural equivalence of the construct.

The step from construct to measurement operation is also made here. The question should be addressed as to whether there is cross-culturally a sufficiently overlapping set of behaviors that can be taken to reflect the construct. If such a set is available, a random sample of such behaviors can be drawn in order to obtain a fair representation of the construct in the instrument. Triandis (1978) has argued that the constructs we intend to measure are often much broader than the behaviors covered by the items. A poor sampling of the domain of behaviors can easily lead to sweeping statements about cross-cultural differences that do not generalize to other measures of the same psychological construct (and low correlations with these other measures).

A further specification of measurement operations is made in the second stage. In this step, method bias is a major threat. The adequacy of stimulus and response formats, test administration procedures, and various other practical aspects regarding the test and its administration have to be scrutinized. If there is reason to assume that one or more aspects could endanger the equivalence, it may be desirable to carry out a pilot study. Thus, the effect of test administrator characteristics such as age and gender, the adequacy of the test instructions, or the clarity of examples in the instruction could be studied. It can be enlightening to administer an instrument in a nonstandard fashion, for instance by asking individuals from the target group to motivate their responses. Such a procedure addresses the question as to whether responses have the intended psychological meaning.

Translation can be troublesome in the case of an instrument that was developed and validated for one language group. It is not uncommon to find that such instruments have items with idiomatic expressions; it may not be easy to find an equivalent in another language that has the same meaning, familiarity, and clarity.

The final stage in test adaptations involves the analysis of item bias. Instrument developers should identify problematic items or aspects of the instrument which may be inadequate to one or more of the intended populations (Hambleton, 1994, p. 238). In addition, they should provide evidence for the validity of the instrument in all target populations.

It may be clear from the description that in test adaptations and, more generally in cross-cultural comparisons, equivalence cannot be established by relying exclusively either on a priori procedures (such as instrument design) or a posteriori considerations (such as item bias analysis). Equivalence is the result of a continuous effort to maintain high standards in all stages of the test development and adaptation process.

Conclusion

Cross-cultural psychology owes its existence to the presence of cultural differences. More than ever before, people nowadays have first-hand experience with such differences. Psychologists can render important services in the area of cross-cultural encounters. Dissemination of the knowledge and experience that cross-cultural psychologists have gained during the last decades, can help to enhance the level of professional services.

In this paper, we focused on assessment in a cross-cultural context. An educational test that is shown to be reliable and valid in the US may show essential flaws in, say, a non-Western context. A personality inventory that is adequate for German natives can be inadequate for migrants in that country. In the future there will be an increasing demand for instruments that can be applied in a multicultural setting. We will have to commit more time and resources to the development and validation of cross-culturally adequate instruments. Score comparisons across cultural groups should not be invalidated by bias.

A study of the adequacy of an instrument in a cross-cultural context should always start from a bias analysis at the construct level which amounts to answering the question as to whether the instrument scores have the same psychological meaning across the cultures studied. The history of psychological testing has shown several examples of strong statements about cross-cultural differences which, upon closer examination, were based on implausible assumptions about the validity of assessment procedures. When empirical bias checks on the accuracy of these procedures are carried out, such overgeneralized statements will become less likely.

The distinction among construct, instrument, and item bias allows bias to be placed in a broad context. During the last few decades bias has become almost synonymous with item bias (or differential item functioning). The development and refinement of item bias techniques has enlarged the tool kit for culture-comparative studies. We concur with the view that item bias techniques should be routinely applied in cross-cultural research, in particular when the study is exploratory and no clear theoretical framework is available to account for cross-cultural differences. However, the emphasis on item bias techniques should not overrate their relevance. Item bias is an important source of problems in intergroup comparisons but particularly when the cultural distance between the groups studied is large, cross-cultural comparisons can also be challenged by undesirable sources of variance that go beyond individual items. We should not underrate the influence of construct and method bias. If during the coming decades the development of models for construct and method bias will be pursued with the same vigor as displayed in the development of item bias techniques, we will undoubtedly witness important methodological innovations in intergroup comparisons.

Author's Address:
Professor Fons J.R. van de Vijver
Department of Social Sciences
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

References

Bentler, P. M. (1992). *EQS structural equation program manual.* Los Angeles: BMDP Statistical Software.

Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting item bias.* Baltimore: The Johns Hopkins University Press.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, pp.389–444). Boston: Allyn & Bacon.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

Endler, N. S., & Magnusson, D. (1976). *Interactional psychology and personality.* New York: Wiley.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: a progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9,* 1–18.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

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

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.

Irvine, S. H., & Carroll, W. K. (1980). Testing and assessment across cultures. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2, pp. 181–244). Boston: Allyn & Bacon.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8.* Chicago: Scientific Software International.

Kendall, I. M., Verster, M. A., & Von Mollendorf, J. W. (1988). Test performance of blacks in South Africa. In

S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context* (pp. 299–339). Cambridge: Cambridge University Press.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105–118.

Ombrédane, A., Robaye, F., & Plumail, H. (1956). Résultats d'une application répétée du matrix-couleur à une population de Noirs Congolais. *Bulletin, Centre d'Etudes et Recherches Psychotechniques*, *6*, 129–147.

Poortinga, Y. H., & Foden, B. I. M. (1975). A comparative study of curiosity in black and white South African students. *Psychologia Africana*, Monograph Supplement, 8.

Poortinga, Y. H., & Malpass, R. S. (1986). Making inferences from cross-cultural data. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 17–46). Beverly Hills, CA: Sage.

Serpell, R. (1979). How specific are perceptual skills? *British Journal of Psychology*, *70*, 365–380.

Serpell, R. (1993). *The significance of schooling. Life-journeys in an African society.* Cambridge: Cambridge University Press.

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

Sternberg, R. J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology*, *49*, 607–627.

Super, C. M. (1981). Behavior development in infancy. In R. H. Munroe, R. L. Munroe, & B. B. Whiting (Eds.), *Handbook of cross-cultural human development* (pp. 181–270). New York: Garland SPTM Press.

Triandis, H. C. (1978). Basic research in the context of applied research in personality and social psychology. *Personality and Social Psychology Bulletin*, *4*, 383–387.

Van de Vijver, F. J. R. (1994). Item bias: Where psychology and methodology meet. In A. Bouvy, F. J. R. Van de Vijver, P. Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 111–126). Lisse: Swets & Zeitlinger.

Van de Vijver, F. J. R., Daal, M., & Van Zonneveld, R. (1986). The trainability of formal thinking: A cross-cultural comparison. *International Journal of Psychology*, *21*, 589–615.

Van de Vijver, F. J. R., & Leung, K. (in press). Methods and data analysis of cross-cultural research. *Handbook of Cross-Cultural Psychology.*

Van de Vijver, F. J. R., & Lonner, W. (in press). A bibliometric analysis of the Journal of Cross-Cultural Psychology. *Journal of Cross-Cultural Psychology.*

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

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. 389–418). New York: American Museum of Natural History.

Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement*, *4*, 141–155.