Towards a Theory of Bias and Equivalence

FONS J. R. VAN DE VIJVER

Bias refers to the presence of nuisance factors in cross-cultural research. Three types of bias are distinguished, depending on whether the nuisance factor is located at the level of the construct (construct bias), the measurement instrument as a whole (method bias) or the items (item bias or differential item functioning). Equivalence refers to the measurement level characteristics that apply to cross-cultural score comparisons; three types of equivalence are defined: construct (identity of constructs across cultures), measurement unit (identity of measurement unit), and scalar equivalence (identity of measurement unit and scale origin). Bias often jeopardizes equivalence. Implications of the occurrence of bias on equivalence are described. Examples of how equivalence can be enhanced in multilingual studies are given.

1. Introduction

Cross-cultural research is a generic name here for all comparative studies that involve either different nation states or different cultural groups within a single country. This kind of research is coming of age. A recent tally of *PsycLit*, an electronic medium publishing summaries of a large number of psychology journals and books, showed that during the last ten years there is a continuous increase of the number of publications dealing with cross-cultural differences (Van de Vijver & Lonner, 1995). Surveys in social sciences will probably reveal the same picture. The increased interest may be related to societal developments. Due to large migration streams, Western countries have become multicultural. For example, in the largest cities in the Netherlands about half of the pupils entering primary school are not native Dutch. In the same vein, it is predicted that by 2020, cities like San Francisco and Los Angeles will have more Hispanic than White Anglo residents. The increased interest may also be fueled by the internationalization of economic life. There are more companies than ever before that operate on an international

market. The booming market of intercultural communication training provides a telling example of this interest.

Although it is reassuring to see the tremendous interest in cross-cultural studies, it is regrettable that there is no generally accepted way of dealing with issues that are specific to cross-cultural research. One can come across empirical studies in which Western instruments have been applied without considering the cultural appropriateness of the measure. There are too many studies in which a test is administered in two cultural groups and in which the only question addressed refers to the difference in average score of the two cultural groups. A comparison of average scores should be preceded by an analysis of the suitability of the instrument. Unless a good theoretical framework is available which can rule out various bias sources, the observation of a significant difference is often open to multiple interpretations such as differential stimulus familiarity (in the case of mental tests) and differential social desirability (on personality and attitude questionnaires). Unfortunately, we do not have well-established and widely adopted practices in cross-cultural search to deal with issues like instrument feasibility and multiple interpretations.

In order to establish such practices we will need to have a theoretical framework that attempts to incorporate aspects that are specific to cross-cultural research. In the present author's view, bias and equivalence are concepts that form the core of such a framework. It will be argued that bias and equivalence are concepts that can guide our plans and actions at all stages of a project, in much the same way as the concepts of validity and reliability underlie many decisions taken in intracultural research. Bias can be viewed as the generic name for all validity-related issues that are specific for cross-cultural research.

In the next section bias and equivalence are defined. The third section links these theoretical concepts to such well-known problems in cross-cultural research as sample incomparability. The fourth section applies this framework to problems encountered in multilingual studies. Conclusions will be drawn in the final section.

2. Bias and Equivalence Defined

The concepts of bias and equivalence have their own history in cross-cultural psychology. *Bias* is related to validity. An instrument is biased if its scores do not have the same psychological meaning across the cultural groups involved; more precisely, an instrument is biased if statements about (similarities and differences of) its scores do not apply in the psychological domain of the scores. For example, individual differences in intelligence test scores may reflect differences in intelligence in a single cultural group, whereas intergroup differences may be largely due to differences in education and test experience. Equivalence has historically become associated with the measurement level at which cross-cultural comparisons can be made. Suppose that in the example of the intelligence test individual differences are measured at ratio level in each cultural group. *Equivalence* refers to the question whether there is any difference in measurement level of within- and between-group comparisons. If the measure is biased against some cultural group, individual differences within a cultural population and across cultural populations are not measured at the same scale.

Three characteristics can be derived from these definitions. First, bias refers to unintended sources of variation that constitute alternative explanations of intergroup differences. If bias is present, cross-cultural score differences are not engendered by the target construct (e.g., intelligence or political affiliation) but by some other characteristic (e.g., social desirability or education). Second, bias and equivalence are not intrinsic to an instrument but characteristics of a specific cross-cultural comparison. Both instrument and sample characteristics will influence the likelihood of occurrence of bias. A questionnaire that can be used to measure political affiliation in, say, France and Germany may be biased in a comparison of France and China. Bias will often increase with the cultural distance to be bridged by the instrument and is also more likely when an instrument shows more cultural saturation. In particular in mental testing much effort has been invested in the development of instruments that can be applied across a wide variety of cultures. Labels used in the past for these tests, such as "culture-free" and "culture-fair"

(e.g., Cattell, 1940; Cattell & Cattell, 1963), sound presumptuous to us; still, the underlying idea that stimulus features can unintentionally and systematically distort observed cross-cultural differences has never been challenged. Finally, bias is a source of systematic variation that is -- at least in principle -- replicable across parallel instruments administered to the same samples.

2.1 Three Types of Bias

Table 1. Types of bias and their description

Type of bias	Description
Construct bias	• dissimilarity of constructs
Method bias	
• Sample bias	• incomparability of samples
• Instrument	• stimulus features that induce cross-cultural differences such as stimulus familiarity
• Administration	• procedural aspects such as communication problems
Item bias	• anomalies at item level such as poor translations

Following Van de Vijver and Leung (1997) three types of bias will be distinguished (cf. Table 1). The first is *construct bias*. It is characterized by dissimilarity of construct across cultures. An example comes from Ho's (1996) work on filial piety in China. The concept refers to the behaviors associated with being a good son or daughter. In Western countries the core of the concept is made up of immaterial aspects such as love and respect; the Chinese concept is broader. In China it is more commonly expected that children play an active role in taking care of their parents once these are unable to support themselves. A

Western-based measure of filial piety will insufficiently cover the Chinese concept while a Chinese questionnaire will be overinclusive according to Western standards; in Embretson's (1983) words, the test will show a poor construct representation. If one is interested in a cross-cultural comparison of constructs that show or are susceptible to construct bias such as filial piety, there is a need to clearly define the behaviors included in the measure.

Method bias is a generic name for all sources of bias emanating from methodological-procedural aspects of a study. The name was coined because in empirical papers most sources of bias meant here are described in the method section. This type of bias can be further subdivided in three subtypes. The first is *sample bias*, subsuming all differences in scores that are related to specific aspects of a sample. Comparability of samples can be a cumbersome issue in cross-cultural comparisons. Two types of sampling schemes are often employed in cross-cultural studies. The first is based on random sampling and aims at securing the results from a single sample to a cultural population at large. The second applies a matched sampling procedure and attempts to control or at least to measure the influence of a potentially confounding variable such as age or education on a target variable. For instance, if one is interested in religious beliefs in different countries, the educational level of the interviewees may be relevant to consider. Sample bias is particularly important to take into account in an examination of culturally highly divergent groups. A random sampling scheme may amount to a comparison of dissimilar groups in terms of background characteristics that are related to instrument scores (e.g., education). On the other hand, a matching procedure may yield atypical samples (e.g., matching Aboriginals and Australians from European descent on education may yield atypical groups in either or both populations). A common way to reduce such sampling problems is the measurement of potentially confounding variables at individual level. In many cases it may be possible to apply statistical procedures to examine the influence of confounding variables such as an analysis of covariance or hierarchical regression procedures (Poortinga & Van de Vijver, 1987).

Instrument bias is the second type of method bias. It is induced by instrument characteristics to which individuals from different cultural groups react in a consistently dissimilar way. Examples are stimulus familiarity (which can influence mental test scores) and differential social desirability or response styles (in personality and attitude measurement). *Administration bias* is triggered by communication problems (e.g., poor mastery of the testing language by one of the parties), interviewer characteristics (e.g., sex and cultural group), or other procedural aspects of the data collection.

Item bias (also known as *Differential Item Functioning*) is the third type of bias. It refers to anomalies of an instrument at item level. Examples are poor translations. Hambleton (1994) gives an example from a Swedish-English comparison of educational achievement: "Where is a bird with webbed feet most likely to live? (a) in the mountains; (b) in the woods; (c) in the sea; (d) in the desert." In the Swedish translation "webbed feet" became "swimming feet," thereby giving a clear cue about the correct answer. Item bias has received much more attention in the literature than construct and method bias. For example, there is a widely accepted, statistically-oriented definition of item bias (e.g., Holland & Wainer, 1993). An item is said to be biased if persons from different cultural groups with the same score on the underlying trait have the same expected score on the item. In other words, persons who are equally dominant (or whatever is measured) and who come from different groups should have the same averages on the item. Equal standing on the underlying trait is usually derived from the total test score.

Numerous techniques have been developed to identify item bias. The most popular technique to date is the Mantel-Haenszel procedure which detects bias in dichotomously scored items (Camilli & Shepard, 1994; Holland & Wainer, 1993). The technique for interval-level scores described here closely follows the rationale of the Mantel-Haenszel procedure. Suppose that a test of dominance consisting of 10 five-point Likert-type items is administered to 400 persons in two countries. An item bias procedure starts with the computation of total test scores (i.e., the sum scores on the 10 items). These range from

10 (10 x 1) to 100 (10 x 10). The extreme scores of 10 and 100 are not taken into account, because by definition persons with these scores have identical response profiles for all

items. The remaining scores are split up into score levels; the number of score levels will be determined by the total sample size; a group size of at least 50 persons in each score group is recommended. An analysis of variance is carried out, with culture and score level group as independent variables and item score as dependent variable. An item is said to be uniformly biased (Mellenbergh, 1982) if the main effect of culture is significant. This implies that for each observed total score level the item is consistently easier or more endorsed in one culture than in another. An item is said to show nonuniform bias if the interaction of score level and culture is significant. In such a case the cross-cultural score differences vary with the observed total test score. In empirical applications, uniform bias is much more common than nonuniform bias.

2.2 Four Types of Equivalence

There is a hierarchical order in the types of equivalence presented here (cf. Table 2). The first refers to the incomparability of constructs across cultures and is labeled *construct inequivalence*; it amounts to "comparing apples and oranges." The other three types show some form of equivalence. The weakest type of equivalence is *construct equivalence*, also known as *functional equivalence* and *structural equivalence*. It occurs when the same

Table 2. Types of equivalence and their description

Type of equivalence	Description
Construct inequivalence	dissimilarity of constructs
Construct equivalence	same construct is measured in each cultural group
Measurement unit equivalence	same scale (measurement unit) with different origins in each cultural group
Scalar equivalence	same scale with same origin in each cultural group

construct has been measured across cultural groups (not necessarily using the same instrument). Construct equivalence is sometimes studied in a comparison of nomological networks across cultures, addressing the question of the construct validity of the measure in each cultural group. Factor analysis is a more frequently employed procedure. In most instances, an exploratory factor analysis is carried out separately in each culture, followed by a target rotation procedure (e.g., Mc Donald, 1985) and the computation of factorial agreement. The target rotation is needed in order to deal with the freedom in rotating factor analytic solutions. So, first the solutions obtained in two cultural groups should be rotated to each other before the agreement can be computed (Van de Vijver and Leung, 1997b, provides an SPSS procedure to carry out the target rotations and compute the agreement index). As an example, Piedmont and Chae (1997) describe the development of a Korean version of a measure of the Big Five personality factors (e.g., McCrae & Costa, 1985), originally developed for the US. In the literature one also finds applications of structural equation modeling to examine construct equivalence. In most cases a confirmatory factor analysis is fitted to the data and the cross-sample stability of the parameters is scrutinized. Taylor and Boeyens (1991), for example, applied confirmatory factor analysis, among other techniques, to study the adequacy of the South African Personality Questionnaire among Blacks and Whites in South Africa.

The third type is *measurement unit equivalence*. We assume here, as below, that the measure is of interval or ratio level in all the cultural populations studied. A measure shows this type of equivalence if the measurement unit is identical across groups while the origins differ. As an example, suppose that temperature is measured using Celsius and Kelvin scales. The measurement units are identical but there is a constant difference (an offset) of 273 degrees of the measures. This type of equivalence will arise if the same instrument has been administered across cultures and method bias (e.g., stimulus familiarity) influences the measure. Individual differences may be measured at ratio level in each group while there is no comparison possible across cultures. Unlike the temperature example, we hardly ever know the offset in measures in the social and behavioral sciences.

In the case of *scalar equivalence* or *full score comparability*, the same interval or ratio level applies to measures in the cultures compared. This is the type of equivalence assumed when averages are compared across cultures, such as in *t* tests and analyses of variance.

3. The Influence of Bias on Equivalence

Bias can be seen as a threat to the validity of cross-cultural studies in that it can lead to inequivalence. The relationship between bias and equivalence is schematically presented in Table 3.

Table 3. Is the level of equivalence affected by bias? (after Van de Vijver & Leung, 1997b)

Type of bias		Level of equivalence		
		Construct	Measurement unit[a]	Scalar[a,b]
Construct bias		yes	yes	yes
Method bias:	uniform	no	no	yes
	nonuniform	no	yes	yes
Item bias:	uniform	no	no	yes
	nonuniform	no	yes	yes

[a]The same measurement unit is assumed in each cultural group;
[b]The same origin is assumed in each cultural group.

There are a few rules underlying the table:

- higher types of equivalence are less robust against bias, for example, scalar equivalence is more susceptible to bias than measurement unit equivalence.
- in terms of actions required for recovery, construct bias is more consequential than are method and item bias;
- nonuniform bias is more consequential than uniform bias because nonuniform bias affects both the origin and the measurement unit of a measure while uniform bias influences merely the origin of the scale.

Scalar equivalence is the strictest type of equivalence, allowing for statements of the type "Culture A has a higher score on propensity F than Culture B." In order to make such strong statements, the absence of any bias is assumed. On the other hand, if one is only interested in the construct equivalence, neither item bias nor method bias will be a threat.

In many empirical applications a choice has to be made whether measurement unit equivalence or scalar equivalence applies. The heated debates about racial differences in intelligence focus on this issue. In the terminology of the present chapter, the debate is about the presence or absence of method bias. In many instances, method bias will lead to an offset in the scales: method bias will induce differences in average scores of cultural groups. Cross-cultural differences in stimulus familiarity, social desirability, and response styles tend to affect many items of an instrument; hence, they will often exert a more or less uniform influence on most or all items of an instrument. From a statistical perspective such an influence may well show up as a significant difference in average scores (e.g., in a *t* test or analysis of variance). Yet, such a cross-cultural difference can be mistakenly interpreted as a real difference on a target construct such as intelligence, while an interpretation in terms of some other characteristic (e.g., educational quality) is more appropriate.

3.1 Example: Multilingual Studies

Multilingual studies are an important area of application of the bias and equivalence issues described above. In most multilingual projects a target instrument is already available that has shown desirable characteristics (reliability and validity) in a particular linguistic group; this instrument is translated for use with other lingustic groups. Studies in which an instrument is simultaneously developed in different languages are less common. Therefore, the present discussion will mainly focus on successive development.

Whereas in the past there has been a tendency to see the linguistic aspects of a translation as the focal area of attention in multilingual studies, there is now a growing awareness that more is involved in the translation of an instrument than rendering text from a source into a target language. In the behavioral sciences, there is rarely much interest in the specific contents of questions and items. Instead, instruments are almost always a means to an end and the operationalizations as expressed in questions and items provide access to underlying constructs, such as political involvement, alienation, and egalitarian commitment. Multilingual studies are often based on the tacit assumption that a careful translation of the instrument will lead to a full transfer of all measurement characteristics such as construct validity and reliability. In the terminology of the present chapter, such a full transfer amounts to an assumption of bias-free measurement and the attainment of the highest level of equivalence possible. The transfer of characteristics from a source-language version to a target language should be empirically scrutinized, since the transfer of the characteristics of the original instrument can be anywhere between absent and complete. In order to maintain the highest level of equivalence possible, the translation and subsequent application of an instrument should be as free of bias as possible. In this, linguistic aspects are important, but not the only ones to be considered. Multilingual studies should focus on validity issues (cf. Bracken & Barona, 1991; Hambleton, 1994; Vallerand, 1989; Van de Vijver & Hambleton, 1996).

In retrospect, it is probably fair to say that the theoretical framework of multilingual studies has become broader in recent times. Recommendations about how to

carry out multilingual studies tended to describe procedures for arriving at accurate translations and provide rules for (in)appropriate item writing, such as the avoidance of the passive and long sentences or the care needed in using referential words such as "his," "her," "this," and "that" because languages differ in their systems of reference. The more recent treatment of multilingual studies from a validity perspective is an acknowledgment of the potential threat of bias and the need to minimize bias in all stages of such a study. A group of researchers recruited from several international psychological associations, headed by Ronald Hambleton (University of Amherst, Massachusetts), recently formulated a set of *guidelines* on how to carry out multilingual studies. Instead of discussing the guidelines (see Hambleton, 1994; Van de Vijver & Hambleton, 1996), I shall briefly present the first two principles which adequately capture the general atmosphere of all guidelines:

> Principle 1. Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.
>
> Principle 2. The amount of overlap in the constructs in the populations of interest should be assessed.

It is characteristic for this approach that central principles of multilingual studies do not relate to linguistic issues but to the reduction of bias and the enhancement of construct validity of the measures.

A multilingual study that is carried out from a validity perspective does not primarily address the question of the translation of an instrument but deals with the question of how to measure the particular construct of the source instrument in the target group, using the characteristics of the latter instrument as much as possible. Such an approach is less direct and more involved than preparing a translation of an instrument; yet it will increase the likelihood that a variety of questions are addressed directly which are answered implicitly, though probably incorrectly, in direct translations, such as:

- Do the items cover the construct in the target group adequately?
- Does the instrument have a format and scoring that is appropriate in the target group?
- Are all items relevant and adequately phrased for the target group?

The broad perspective adopted by validity studies has various implications. A literal translation, quite often seen as the only available option in multilingual studies, is one of the possibilities from a validity perspective. In general, translation studies can apply three strategies depending on the type of bias to be expected. First, when construct bias can be expected to threaten a literal translation of the original measure, the *assembly* of an entirely new instrument may be needed to obtain a good representation of the construct in the new cultural context. A good example can be found in the work by Cheung et al. (1996). These authors argued that Western personality measures do not address all relevant dimensions of the Chinese personality. They developed the Chinese Personality Assessment Inventory. In order to examine construct bias of common Western measures, a pilot study was carried out addressing important characteristics of personality as seen by Chinese subjects. The pilot study pointed to the need to include constructs such as "face" and "harmony." The final version of the inventory has both universal and culture-specific aspects of personality. Their study illustrates various features of an assembly approach towards test development: adequate representation of a local construct instead of cross-cultural comparability (and scalar equivalence) is the aim of the project, thereby maximizing the suitability of the instrument for the local context though precluding the opportunity to compare scores across cultures. Furthermore, assembly studies tend to require huge amounts of resources (time and money).

Adaptations constitute the second type of multilingual study. Some (or even most) stimuli are considered appropriate but as a whole the instrument is not taken to yield an appropriate measure of the target construct. Adaptations amount to the literal translation of some stimuli and, depending on the specific features of the instrument, to adding, changing, or removing other stimuli. Adaptation will be the preferred choice when there

is an incomplete overlap in the behaviors or attitudes associated with a construct. A good example of the adaptation option is the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970). This instrument had been translated into more than 40 languages. Most versions are not literal translations of the English-language original, but are adapted in such a way that the underlying constructs, state and trait anxiety, are measured adequately in each language (e.g., Laux, Glanzmann, Schaffner, & Spielberger, 1981). Another example is the Minnesota Multiphasic Personality Inventory (Dahlstrom, Welsh and Dahlstrom, 1972), which has been adapted to various cultural contexts. The constructs of the tests are broad and various items have a limited applicability outside the US, where the inventory was developed. A Mexican adaptation has been described by Lucio, Reyes-Lagunes, and Scott (1994) and a Chinese adaptation by Cheung (1989).

The statistical analyses of adapted instruments often amount to an examination of the construct validity of the new instrument. For example, Cheung (1989), who adapted the MMPI to China, provides evidence for the validity of the scale by examining its ability to discriminate between normals and patients and by computing profiles for different diagnostic groups. She reported patterns similar to those found in the US.

Due to developments in statistical methods, the opportunities for analysis have been expanded in the last decades. The first important development is item response theory (e.g., Hambleton, & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Molenaar & Fischer, 1995). Scores of subjects can be compared across instruments that are based on partially dissimilar item sets. As a hypothetical example, suppose that a German inventory of 15 items to measure political interest is translated for use in an entirely different political system. Furthermore, let us assume that five items have to be replaced by new items, leaving a common set of 10 items. If the assumptions of item response theory are met, a comparison of scores and even a statistical comparison of means of cultural groups in a *t* test can be obtained. The most relevant assumption will be that the 15 items measure a single latent trait in both groups and that the 10 common items measure the same latent trait in both groups. Statistical tests of the assumptions are

available (Hambleton, & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Molenaar & Fischer, 1995).

The second relevant development has taken place in the area of factor analyses, both exploratory (e.g., Kiers, 1990; Kiers & Ten Berge, 1989) and confirmatory (e.g., Bollen, 1989; Bollen & Long, 1993; Byrne, 1989, 1994). Target rotations are a common way to explore the similarity of factors obtained in an exploratory factor analysis across cultural populations (cf. van de Vijver & Leung, 1997a,b). Because factor analytic solutions can be arbitrarily rotated, solutions obtained in different populations have first to be rotated towards each other (i.e., their agreement has to be maximized) before their correspondence can be assessed. The computation of an agreement of adapted instruments amounts to a factor analysis of all items in each cultural group, thereby allowing that both common and culture-specific items define factors, and a target rotation of the common items. (The culture-specific items are defined as missing values). Van de Vijver and Leung (1997**??**) provide an SPSS procedure to carry out target rotations and compute agreement indices (including the frequently reported Tucker's phi).

The way in which so-called multisample analyses in confirmatory factor analysis deal with test adaptations is somewhat similar. Both common and culture-specific items are utilized to get an adequate factorial representation in each cultural population. The use of multisample procedures in confirmatory factor analysis allows for a fine-grained (i.e., item-level) test of similarities of loadings of common variables across cultural populations. As an example, De Groot, Koot, and Verhulst (1994) examined the cross-cultural stability of the Child Behavior Checklist, a measure of child pathology, in the US and the Netherlands. Most syndromes (factors) were similar across these countries.

Both item response theory and structural equation modeling have enlarged the tools of the cross-cultural researcher in an interesting way; however, the limitations of the techniques can be easily overlooked. Suppose that in our example there were five common and ten culture-specific items. With such a small core of common items, the common and ten culture-specific items. With such a small core of common items, the

culture-specific aspects may describe salient aspects of the construct not covered by the other items; the common core may underrepresent the construct. The poorer the representation will be, the more likely it will become that construct bias endangers the comparability of scores.

By far the most popular option in multilingual studies is *application*. It amounts to the literal translation of the original stimulus material. Translation-backtranslations are often employed to arrive at appropriate translations of stimulus material. In most cases the translator will be hired for his or her linguistic expertise. Such an approach may be inappropriate if method or construct bias jeopardize the equivalence of scores. A so-called committee approach in which persons from different areas of expertise participate is better equipped to deal with the complexities of method and in particular construct bias (cf. Hambleton, 1994).

The literature contains many examples of the application option. Smith, Tisak, Bauman, and Green (1991) studied the equivalence of a translated circadian rhythm questionnaire in English and Japanese. Several discrepancies between the original and translated scales were found. Ellis, Becker, and Kimmel (1993) studied the equivalence of an English-language version of the Trier Personality Inventory and the original German version. Among the 120 items tested, 11 items were found to be biased.

The reason for the popularity of literal translations can be easily appreciated. Compared to the assembly of new instruments or the adaptations of existing ones, applications are cheap and retain all opportunities for scalar equivalence. As can be expected, these advantages are not without costs: applications require the absence of bias. Reading the cross-cultural literature, one cannot escape from the impression that the assumption of the absence of bias is often readily made and that claims about absence of bias are only infrequently substantiated. In the social and behavioral sciences, we are often inclined to work from the assumption that our measures are unobtrusive (Webb, Campbell, & Schwartz, 1966), despite the impressive evidence to the contrary. Thus, in a recently completed meta-analysis of cross-cultural differences in cognitive test

performance, the present author found that commercially available Western tests such as Raven's Colored, Standard, and Advanced Progessive Matrices and the Wechlser intelligence scales for children and for adults yielded consistently larger cross-cultural differences than did locally developed non-Western tests (Van de Vijver, 1997).

Validity Enhancement in Multilingual Studies

Many multilingual studies are designed with the aim to compare scores or score patterns across languages. Such an aim amounts to the attainment of the highest level of measurement equivalence possible. Various measures can be taken to enhance the validity of multilingual studies (cf. Van de Vijver & Tanzer, 1997). Obviously, a listing of the measures cannot be exhaustive and some selection criterion is needed. The present overview provides a small overview of frequently proposed measures. The types of bias that were distinguished previously (construct, method, and item bias) constitute the framework in which the measures will be presented (see Table 4).

There are a few ways in which construct bias can be adequately addressed. In the first, decentering (Werner & Campbell, 1970), an instrument is simultaneously developed in all target languages. Ideally, a team with an expertise in both psychology and linguistics is set up for each language. These teams exchange information about the construct and its associated behaviors or attitudes. Culture-specific aspects, such as problematic wording or the use of particular answer rubrics, are likely to be detected and can be removed. An instrument developed this way will not have the implicit or explicit references to the cultural background of the test developer that are characteristic for many measures in the social and behavioral sciences. An interesting variation to this technique is the so-called 'convergence approach,' in which researchers and cultures are crossed. As an example, an Indian and a German political scientist want to study political interest. Both write an inventory for their own cultural group. The instrument is translated in the other language. Both instruments are then administered in both countries. A comparison

Table 4. Strategies for Identifying and Dealing with Bias in Cross-Cultural Assessment (from Van de Vijver & Tanzer, in press)

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) • nonstandard instrument administration (e.g., "thinking aloud") • cross-cultural comparison of nomological networks (e.g., convergent/discriminant 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 exercises) • 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 • detailed manual/protocol for administration, scoring, and interpretation • 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

of the results may provide insight into the universal and culture-specific aspects of the instrument.

Another set of measures addresses construct and/or method bias. Examples are the use of bilingual subjects and of local surveys. If there is doubt about the applicability of an instrument, nonstandard administrations (e.g., think aloud protocols) can be an aid in the identification of problematic aspects. Another way of addressing construct and/or method bias is the cross-cultural comparison of nomological networks. Such a comparison attempts to answer the question whether an instrument shows a convergent and discriminant validity that may be expected in each culture. Structural equation modeling provides a data-analytic tool to compare nomological networks across cultures.

The measures that can be taken to reduce method bias are numerous. The general procedure behind most measures is the reduction or measurement of relevant confounding variables. Examples aimed at the reduction of nuisance factors are the extensive training of test administrators/interviewers and the preparation of a detailed protocol for administering, scoring, and interpreting an instrument. When the cultural distances to be bridged by an instrument are large, procedures to reduce the influence of confounding factors may be insufficient. For example, when groups of literate and illiterates are compared, lengthy instructions and well-defined administration guidelines cannot make up for the immense differences in relevant background variables. In such cases, an alternative to reduction may be measurement of the most relevant background variables. The influence of these variables can be assessed in an analysis of covariance or hierarchical regression analysis.

An interesting way to examine method bias is the repeated administration of the same instrument in various cultural groups and the examination of score changes, usually score increments, upon retesting. If subjects with similar scores on the pretest show differential gain patterns, strong evidence for method bias has been obtained. Gain patterns on cognitive tests that are larger in non-Western groups than in Western groups

have been reported (e.g., Kendall, Verster, & Von Mollendorf, 1988). Nkaya, Huteau, and Bonnet (1994) administered Raven's Standard Matrices three times to sixth graders in France and Congo. Under power conditions (i.e., when no time limit was applied) a moderate improvement from the first to the second and no progress from the second to the third administration were observed in both groups. Under timed conditions both groups progressed rapidly from the first to the second; however, only the Congolese pupils progressed from the second to the third session. Such findings retrospectively cast doubt on the score equivalence of the first administration.

Disturbances at item level are commonly detected by either of two procedures. The first is the use of judgmental procedures. A few years ago a committee of Dutch psychologists carried out a content analysis of commonly employed psychological tests; the adequacy of these instruments for individuals whose native tongue is other than Dutch was judged. The committee concluded that ethnocentrism is rampant (Hofstee, 1990). The second procedure to detect item-level disturbances is the use of item bias techniques (which have been described before).

Despite their relevance and widespread use, particularly in the area of educational testing, these techniques are not without their problems. Apart from statistical-technical problems mentioned earlier (such as the need for huge samples), there is a problem of interpretation: expert judgments and item bias procedures are more or less consistently found to be unrelated. Sources of item anomalies as identified by experts such as implicit ethnocentrism are often not flagged as biased by statistical procedures. A recent example is a study by Van Leest (1997) investigating the suitability of two personality questionnaires frequently employed among native Dutch for the selection of migrants in the Netherlands. Experts from minority groups (from the target groups of the study) were asked to judge the instruments. Entirely in line with the Hofstee committee, they found many items inadequate for use among migrants. Statistical procedures also identified many biased items; yet, there was no relationship between the conclusions of the

judgmental and statistical procedures. Furthermore, empirical research has shows that item bias is poorly understood. Item bias is often not at all stable across instruments and samples. Thus, Scheuneman (1987) studied bias in items for American Blacks and Whites on the Graduate Record Examination General Test. Various hypotheses about the influence of formal characteristics on item bias were tested (such as a negative phrasing of item stem, clarity of content, and ordinal position of the correct alternative). Some systematic relationships were found; however, Scheunemann concluded "what emerges most clearly from the study is how little we know about the mechanisms that produce differential performance between black and white examinees" (p. 117). Or in Linn's (1993; 359) words: "The majority of items with large DIF values seem to defy explanation of the kind that can lead to more general principles of sound test development practice".

4. Conclusion

Bias and equivalence are integral elements of each and every cross-cultural study. Bias refers to the absence or presence of nuisance factors while equivalence refers to the implications of bias on the cross-cultural score comparisons to be made. In order to safeguard the highest possible level of equivalence, bias should be scrutinized in each and every stage of an empirical project. Hopefully, a serious concern for bias and equivalence will become a routine consideration in cross-cultural studies, in much the same way as validity and reliability have become standard concepts that have deeply influenced our thinking about to design, administer, score, and interpret test scores. In an era in which cross-cultural encounters are becoming more frequent and cross-cultural research is gaining momentum, it is important to design agreed-upon procedures to carry out such research.

References

Bollen, K.J. (1989): *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K.J. & Long, J.S. (eds.) (1993): *Testing Structural Equation Models*. Newbury Park, CA: Sage.

Bracken, B.A. & Barona, A. (1991): State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International 12*: 119-132.

Byrne, B.M. (1989): *A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analytic Models*. New York: Springer.

Byrne, B.M. (1994): *Structural Equation Modelling with EQS and EQS/Windows: Basic Concepts, Applications, and Programming*. Thousand Oaks, CA: Sage.

Camilli, G. & Shepard, L.N. (1994): *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.

Cattell, R.B. (1940): A culture-free intelligence test, I. *Journal of Educational Psychology 31*: 176-199.

Cattell, R.B. & Cattell, A.K.S. (1963): *Culture Fair Intelligence Test*. Champaign, IL: Institute for Personality and Ability Testing.

Cheung, F.M., Leung, K., Fan, R.M., Song, W.Z., Zhang, J.X. & Chang, J.P. (1996): Development of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology 27*: 181-199.

Dahlstrom, W.G., Welsh, G.S. & Dahlstrom, L.E. (1972): *An MMPI Handbook*. Minneapolis: University of Minnesota Press.

De Groot, A. Koot, H.M. & Verhulst, F.C. (1994): Cross-cultural generalizability of the Child Behavior Checklist cross-informant syndromes. *Psychological Assessment 6*: 225-230.

Ellis, B.B., Becker, P. & Kimmel, H.D. (1993): An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology 24*: 133-148.

Embretson, S.E. (1983): Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin 93*: 179-197.

Hambleton, R.K. (1994): Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment (Bulletin of the International Test Commission) 10*: 229-244.

Hambleton, R.K. & Swaminathan, H. (1985): *Item Response Theory: Principles and Applications*. Dordrecht: Kluwer.

Hambleton, R.K. Swaminathan, H. & Rogers, H.J. (1991): *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

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

Hofstee, W.K.B. (1990). Toepasbaarheid van psychologische tests bij allochtonen. *De Psycholoog 25*: 291-294.

Holland, P.W. & Wainer, H. (eds.) (1993). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Kendall, I.M., Verster, M.A. & Von Mollendorf, J.W. (1988): Test performance of blacks in South Africa. In: Irvine, S.H. & Berry J.W. (eds.), *Human abilities in cultural context* (pp. 299-339). Cambridge: Cambridge University Press.

Kiers, H.A.L. (1990): *SCA: A Program for Simultaneous Components Analysis*. Groningen: IEC ProGamma.

Kiers, H.A.L. & Ten Berge, J.M.F. (1989): Alternating Least Squares Algorithms for Simultaneous Components Analysis with equal component weight matrices for all populations. *Psychometrika 54*: 467-473.

Laux, L., Glanzmann, P., Schaffner, P. & Spielberger, C.D. (1981): *Das State-Trait Angstinventar. Theoretische Grundlagen und Handanweisung* [The German Adaptation of the State-Trait Anxiety Inventory. Theoretical Background and Manual]. Weinheim, Germany: Beltz Test.

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

Lucio, E., Reyes-Lagunes, I. & Scott, R.L. (1994): MMPI-2 for Mexico: Translation and adaptation. *Journal of Personality Assessment 63*: 105-116.

McCrae, R.R. & Costa, P.T. (1985): Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology 49,* 710-721.

McDonald, R.P. (1985): *Factor Analysis and Related Methods*. Hillsdale, NJ: Erlbaum.

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

Molenaar, I.W. & Fischer, G.H. (eds.) (1995): *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer.

Nkaya, H.N., Huteau, M. & Bonnet, J. (1994): Retest effect on cognitive performance on the Raven-38 Matrices in France and in the Congo. *Perceptual and Motor Skills 78*: 503-510.

Piedmont, R.L. & Chae, J-H. (1997): Cross-cultural generalizability of the five-factor model of personality: Development and validation of the NEO PI-R for Koreans. *Journal of Cross-Cultural Psychology 28*: 131-155.

Poortinga, Y.H. & Van de Vijver, F.J.R. (1987): Explaining cross-cultural differences: Bias analysis and beyond. *Journal of Cross-Cultural Psychology 18*: 259-282.

Scheuneman, J.D. (1987): An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement 24*: 97-118.

Smith, C.S., Tisak, J., Bauman, T. & Green, E. (1991): Psychometric equivalence of a translated circadian rhythm questionnaire: Implications for between- and within-population assessments. *Journal of Applied Psychology 76*: 628-636.

Spielberger, C.D., Gorsuch, R.L. & Lushene, R.E. (1970): *Manual for the State-Trait Anxiety Inventory ("Self-Evaluation Questionnaire")*. Palo Alto, CA: Consulting Psychologists Press.

Taylor, T.R. & Boeyens, J.C. (1991): The comparability of the scores of Blacks and Whites on the South African Personality Questionnaire: An exploratory study. *South-African Journal of Psychology 21*: 1-11.

Vallerand, R.J. (1989): Vers une methodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue francaise. *Canadian Psychology 30*: 662-680.

Van de Vijver, F.J.R. (1997): Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology 28*, 670-709.

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

Van de Vijver, F.J.R. & Leung, K. (1997a): Methods and data analysis of comparative research. In: Berry, J.W., Poortinga, Y.H. & Pandey, J. (eds.), *Handbook of Cross-Cultural Psychology* (2nd ed., vol. 1, pp. 257-300). Boston: Allyn & Bacon.

Van de Vijver, F.J.R. & Leung, K. (1997b): *Methods and Data Analysis for Cross-Cultural Research*. Newbury Park, CA: Sage.

Van de Vijver, F.J.R. & Lonner, W. (1995): A bibliometric analysis of the Journal of Cross-Cultural Psychology. *Journal of Cross-Cultural Psychology 26*: 591-602.

Van de Vijver, F.J.R. & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*. (in press)

Van Leest, P.F. (1997): Bias and equivalence research in the Netherlands. *European Review of Applied Psychology*. (in press)

Webb, E.J., Campbell, D.T. & Schwartz, R.D. (1966): *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.

Werner, O. & Campbell, D.T. (1970): Translating, working through interpreters, and the problem of decentering. In: Naroll, R. & Cohen, R. (eds.), *A Handbook of Cultural Anthropology* (pp. 398-419). New York: American Museum of Natural History.