Equivalence

Johnny R. J. Fontaine
Ghent University, Ghent, Belgium

Glossary

construct bias Generic term used to refer to the cultural specificity of the theoretical variable and/or domain underrepresentation.
construct equivalence Generic term used to refer to functional and/or structural equivalence.
cultural specificity of the theoretical variable Occurs when a theoretical variable can be used validly only within a specific cultural context.
domain underrepresentation Occurs when important aspects of the domain that a theoretical variable is assumed to account for are not represented in the measurement instrument.
full score equivalence Occurs when scores that can be directly compared between cultural groups.
functional equivalence Occurs when the same theoretical variable accounts for measurement outcomes across cultural groups.
item bias Occurs when scores on a specific item cannot be compared across cultural groups.
method bias Occurs when method factors have a differential impact on measurements across cultural groups, leading to noncomparability of scores.
metric equivalence Occurs when relative comparisons, for instance, between experimental conditions, are valid between cultural groups.
structural equivalence Occurs when the same measurement instrument forms a valid and sufficient indicator of a theoretical variable across cultural groups.

Introduction

Intensified intercultural exchanges and mass migrations leading to multicultural societies have also influenced the behavioral sciences. These sciences have become more cross-culturally oriented and the types of societal and individual problems with which practitioners are confronted increasingly require attention for cultural diversity. It is against this background that the question to which extent measurements are comparable across cultural groups has become relevant from both theoretical and applied perspectives.

Measuring across cultural groups adds a major complexity to behavioral science measurement. In 1970, Przeworksi and Teune expressed this as follows: "For a specific observation a belch is a belch and nepotism is nepotism. But within an inferential framework, a belch is an 'insult' or a 'compliment' and nepotism is 'corruption' or 'responsibility.'" Identical questions can have different meanings across cultural groups. In order to deal conceptually with comparability or noncomparability of data, the two twin concepts of equivalence and bias have been developed. Although both concepts have a more or less opposite meaning—with equivalence pointing to comparability and bias to noncomparability—historically they have somewhat different roots. Equivalence mostly refers to the question whether scores can be compared, whereas bias mostly refers to causes leading to distortion of comparability.

Since the conceptualization of bias and equivalence is embedded in and builds on general concepts developed for behavioral science methodology and measurement, an overview of these concepts is presented here followed by a discussion of the major levels of equivalence. Then, various factors that can bias cross-cultural measurement are examined followed by a discussion of commonly applied research and data-analytic methods that are used to justify cross-cultural equivalence of data or to detect bias.

Equivalence and Bias: A General Framework

Measurement within the behavioral sciences can be conceptualized as an interplay between three types of

variables, namely, observed, latent, and theoretical variables. The term "observed variables" refers to the concretely observed test behaviors, such as the correct or incorrect answers to a set of items in an inductive reasoning test. According to some implicit or explicit rule, the observed test behaviors are transformed into an estimate of an underlying characteristic, called the latent variable. For example, by coding any wrong answer as zero and any correct answer as 1 and taking the sum of the correct answers, the position of a person on the latent variable can be estimated. The name "latent" variable is used since it is not directly observed; one observes only the correct and incorrect answers. Subsequently, estimated scores on the latent variable are interpreted in terms of a theoretical variable. In a test of inductive reasoning, the test scores are interpreted as referring to an ability for inductive reasoning.

The adequacy of inferences in terms of a theoretical variable based on latent variable score estimates derived from observed variables forms the subject of analysis in the validity framework. The validity of the inferences depends on the links that can be demonstrated between each of the three types of variables. It should be noted that it is not the measurement instrument as such but the behaviors elicited by it within a specific measurement context that determine the validity. Even if prior knowledge or research has established the relevance and representativeness of the items used in an instrument, it is always possible that the items do not elicit the intended behavior. For instance, due to a complex wording of items in an inductive reasoning test, the test may tap verbal abilities rather than inductive reasoning. Moreover, the same stimuli can elicit different behavior in different contexts. For instance, social desirability is likely to have a larger impact on an attitude questionnaire used in a selection context than in a survey context where anonymity is guaranteed. Thus, the adequacy of the inferences derived from an instrument must be demonstrated separately for the various contexts, including cultural contexts, in which that instrument is applied.

Relevance and Representativeness

The items or stimuli of an instrument usually form a limited subset of the domain of phenomena that the theoretical variable is assumed to account for. The aim is to generalize the results of this small subset to an entire domain. Thus, the score on an inductive reasoning test will be generalized to other situations, beyond the testing situation, where inductive reasoning is assumed to play a role. For this generalization to be justified, the observed variables should be both relevant to and representative of the entire domain of phenomena. Irrelevant items would introduce some other theoretical domain(s) into the measurement and lead to systematic noise. If items of an inductive reasoning test contain unfamiliar words, verbal abilities have a systematic, additional impact on the measurement. Moreover, an instrument cannot give any information about aspects of the domain that are not represented in the items. For example, the results of an inductive reasoning test do not give an indication for intellectual functioning in general, since the latter theoretical variable refers to a much broader domain than represented in the test. Although evidence for the relevance and representativeness of the stimulus material forms a necessary condition for a valid interpretation of the test scores, it does not form a sufficient condition; other unintended factors, such as response styles, may interfere. As already mentioned, it must be demonstrated that the selected stimuli elicit the intended behavior in the specific contexts where they are applied.

Psychometric Modeling

Analysis of the relationships between observed and latent variables provides a first possibility for investigating whether the intended psychological phenomena are elicited by a set of stimuli. In psychometrics (but also in sociometrics and in econometrics), mathematical and statistical models have been developed that relate observed item behavior to a position on a latent variable. By specifying precise relationships between items and the latent variable, a psychometric model also makes predictions about how the scores on the items should interrelate. Thus, it can be tested whether a psychometric model adequately accounts for the observed interrelationships. If this is the case, the psychometric model allows estimates of the position of an individual test taker on the latent variable.

The most common psychometric models use two basic parameters to describe the relationship between an observed and a latent variable, namely, an association or weight parameter and a level or intercept parameter. Since these parameters play an important role further on in this article when the levels of equivalence and types of bias are distinguished, these two parameters are presented in somewhat more detail for two prototypical psychometric models, namely, confirmatory factor analysis and the two-parameter logistic model.

Confirmatory factor analysis (CFA) is often used with the measurement of attitudes and personality traits. In this model, it is assumed that both the observed and the latent variables are measured at interval level. The relationship between observed and latent variable can be represented by a regression model, $X_{ij} = a_i + b_i Y_j$, where X_{ij} is the expected score on the observed variable i and Y_j is the position on the latent variable for subject j. The association parameter, or weight, b_i, indicates the expected change in the observed score when there is one unit change in the latent variable. The stronger the

association between observed and latent variable, the higher is the weight. The level parameter or intercept, a_i, is a constant corresponding to the expected score on the observed variable when the score on the latent variable is zero. The higher the intercept of an item, the higher the item scores across positions on the latent variable.

A quite different model is the two-parameter logistic model. This model is used in achievement and ability testing where the observed variables can be scored dichotomously (correct or incorrect) and the latent variable is at interval level. It models the probability of success on a specific item $[P_i(\theta)]$ as a logistic function of the position on the latent variable, namely, $P_i(\theta) = \exp[x_i(\theta - y_i)]/1 + \exp[x_i(\theta - y_i)]$, with y_i representing the item difficulty and x_i the item discrimination. The item difficulty corresponds to the position on the latent trait where the probability of success is 50%. The item discrimination corresponds with the steepness of the logistic curve. It indicates how well correctly responding to the item discriminates between subjects situated below and above the item difficulty level. Although this model may look very different from the CFA model, the natural logarithm of the odds of a correct versus an incorrect answer to an item can be described by the same regression equation that was presented for CFA, namely, $\ln[P_i(\theta)/1 - P_i(\theta)] = a_i + b_i\theta$. In this equation, the weight b_i is equal to x_i and the intercept a_i is equal to $-y_i x_i$. The weight and the intercept parameter have the same interpretation as the association and level parameters in the CFA model.

Nomological Network

A fundamental question in scientific research concerns the interpretation of latent variable scores in terms of the intended theoretical variable, while excluding other theoretical variables. An important source for the meaning of a theoretical variable is the network of predicted positive, zero, and negative relationships with other theoretical variables, called the nomological net. It reflects major information on the scientific theory about that variable. The interpretation of a measurement gains in credibility to the extent that this network of relationships can be empirically confirmed. For instance, the credibility of a score interpretation in terms of inductive reasoning increases if the measurement relates in the theoretically predicted way to other measures, such as tests of deductive reasoning and school success or failure.

Multitrait–Multimethod Approach

One of the major practical concerns in measurement lies in controlling for the possible impact of the specific method that is used on the estimates of a theoretical variable. For example, working with Likert scales can introduce an acquiescent response style. Method variance can be detected by varying the method that is used to measure a theoretical variable. In the validity literature, a multitrait–multimethod research paradigm has been proposed in which each of several theoretical variables is measured by more than one method. By comparing relationships between the measurements of the same theoretical variable using different methods with relationships between the measurements of different theoretical variables using the same method, it is possible to disentangle the variance introduced by a specific method from valid variance that is accounted for by the theoretical variable.

Levels of Equivalence

When the measurement context is extended to more cultural groups, three basic questions can be asked within the general measurement framework presented, namely, whether across these groups (1) the same theoretical variables can account for test behavior, (2) the same observed variables can be used for measurement, and (3) comparisons can be made based on the same latent variables. Since two basic parameters can be distinguished in the relationship between observed and latent variables, these questions lead to four different levels of equivalence. These are as follows: (1) functional equivalence, which implies that a theoretical variable has the same psychological meaning across the cultural groups; (2) structural equivalence, which implies that an observed variable refers to the same latent variable, or—in measurement terms—that the weight parameter linking observed and latent variables differs (nontrivially) from zero in each of the groups; (3) metric equivalence, which implies that the weight parameter between an observed and a latent variable has the same value in the cultural groups and thus that cross-cultural comparisons of score patterns can be made; and (4) full score equivalence, which implies the same value for both the weight and the intercept parameters between observed and latent variables across the groups. This allows for cross-cultural comparisons of scores at face value. Note that these four types are hierarchically ordered in the sense that a higher level of equivalence requires that the conditions for the lower levels are met. Each of these four levels of equivalence is presented in greater detail here (see also Table I).

Functional Equivalence

The first and most fundamental question is whether the same explanatory concept can account for (test) behavior across cultural groups. A valid measurement instrument has to be constructed within each of the cultural groups satisfying the general methodological and measurement requirements as presented previously. Since this level of

Table I Four Levels of Equivalence, with Corresponding Types of Bias and Research and Data Analytic Methods

Level of equivalence	*Answer to question*	*Major conditions*	*Types of bias*	*Type of research and data analytic methods*
(1) Functional equivalence	Can the same theoretical variable account for test behavior across cultural groups?	Similar network of convergent and discriminant relationships with other theoretical variables across cultural groups	(1) Cultural specificity of the theoretical variable	(1) Analysis of nomological network and context variables
(2) Structural equivalence	Can the same observed variables be used across cultural groups?	Stimuli should be relevant and representative for the content domain across cultural groups	(2) Domain under-representation	(2) Analysis of domain via expert judgments or qualitative research
		Stimuli should have a non-trivial weight parameter across cultural groups (= same internal structure)	(3) Method bias • Instrument bias • Administration bias • Sample bias	(3) Multitrait–multimethod measurements
(3) Metric equivalence	Can patterns of scores be compared between cultural groups?	Identical weight parameters across cultural groups	(4) Item bias • Non uniform item bias	(4) Psychometric models for studying relationships between observed and latent variables
(4) Full score equivalence	Can scores be directly compared between cultural groups?	Identical intercept parameters across cultural groups	• Uniform item bias	

equivalence does not require that measurement procedures, and thus observed variables, are the same in each cultural group, only patterns of convergent and discriminant relationships with other theoretical variables can be compared across cultural groups. This means that the measurement should demonstrate the same functional interactions in a network of convergent and discriminant relationships with other variables. Hence, cross-cultural comparability for this level of equivalence is called functional equivalence.

An example of functional equivalence can be found in the work of Ekman and Friesen. For their investigation of emotion recognition in facial expressions among the illiterate Fore in East New Guinea, they could not rely on measurement procedures used in literate societies. Literate respondents were asked to select the emotion term from a written list of terms that best described the emotion displayed in each of a series of photographs of facial expressions. Fore respondents were told three emotional stories and were then asked to mention which story corresponded best with the facial expression. Also, Fore participants were asked to display the facial expression that would correspond to the emotion they would feel in various emotional situations presented by the researchers. Although the measurement procedures were different for the Fore, it can be considered functionally equivalent with the procedure with literate samples.

Structural Equivalence

The second question is whether the same theoretical variable can be operationalized by the same observed variables across cultural groups. Since such observed variables reflect the reactions of respondents to specific stimuli, two conditions that are often treated separately in the literature must both be met. First, the items or stimuli of the instrument should be relevant to and representative of the content domain within each of the cultural groups. Second, it should be demonstrated that the items or stimuli indeed elicit the intended behavior in each of the cultural groups. This implies that each item response should refer to the same latent variable in each of the cultural groups. In terms of psychometric modeling, this means that the latent variable should have positive (nontrivial) weight parameters for each of the observed variables in each of the cultural groups. No further restrictions are imposed on the weight and intercept parameters. Since the relationships between observed and latent variables are referred to as the "structure" of a test, cross-cultural comparability at this level is called structural

equivalence. For instance, in 1995, Schwartz and Sagiv found that 44 of the 56 value items of the Schwartz Value Survey shared a highly stable position in a two-dimensional representation across cultural groups based on the rank order of their mutual conflicts and compatibilities. Although the analyses gave no information about the exact size of weight and intercept parameters (they were based on rank orders), the stable positions pointed to cross-culturally stable nontrivial weights of these 44 items with respect to the underlying dimensions.

In the literature, structural equivalence and functional equivalence are often discussed together as construct equivalence. They both imply the use of the same theoretical variables across cultural groups. However, empirical support for these two levels of equivalence is insufficient to justify quantitative comparisons of scores or patterns of scores between cultural groups.

Metric Equivalence

The third question is whether it is possible to make quantitative comparisons between cultural groups on the basis of the latent variable. The type of comparisons that can be made depends on the restrictions that hold on the intercept and weight parameter in the psychometric model. If only the values of the weight parameter are the same across cultural groups, then it is possible to compare patterns of scores between cultural groups. Equal weight parameters imply that a change in the latent variable leads to the same expected change in the observed variables in each of the cultural groups. The restriction of equal weight parameters implies that the observed and latent variables are measured on the same metric scale across cultural groups. Since the origin of the scale can still differ across cultural groups, only patterns of scores, referring to, for instance, experimental conditions or subgroups, can be directly compared across cultural groups. This level of equivalence is called metric equivalence or measurement unit equivalence in the literature. It can be compared with measuring temperature on a Celsius scale in one cultural group and on a Kelvin scale in another cultural group. The temperatures cannot be directly compared between the two groups; however, relative differences can be compared, such as the difference between the average temperature in summer and in winter.

In a study on information processing comparing Kpelle (Liberian) with U.S. children, metric equivalence was assumed. Children were asked to report the number of dots they had seen on short-term (0.25 s) displays. In one condition, the array of dots was random; in the other, there was patterning. According to the authors, the observed differences in average achievement between Kpelle and U.S. children could have been caused by factors such as motivation and familiarity with the testing procedure. However, it was assumed that these would affect both conditions to the same extent. Only the relative differences between the conditions were compared between the two cultural groups and interpreted in terms of information processing.

Full Score Equivalence

The final and last question is whether scores on a measurement instrument can be compared directly (i.e., at face value) across cultural groups. In addition to the requirements for functional, structural, and metric equivalence, this level of equivalence requires equal intercepts across cultural groups in the regression functions linking the observed and latent variables. This means that if two persons have the same position on the latent variable, exactly the same observed score is to be expected independent of cultural membership. This level of equivalence is therefore called full-score equivalence. Differences between cultural or ethnic groups on a test can be validly interpreted only if there is full score equivalence. In the literature, the term "impact" has been coined to refer to such differences.

Authors who take the stance that differences between ethnic groups on intelligence tests reflect genuine and even genetic differences in intelligence assume full-score equivalence for the instruments measuring intelligence. However, there is a vigorous debate as to whether such an assumption is justified. The next section focuses on possible causes of distortion of equivalence.

Types of Bias

Complementary to the three basic questions for equivalence, there are three basic questions for bias: across cultural groups (1) what can cause bias in theoretical variables, (2) what can cause bias in the observed variables, and (3) what can cause bias in the direct comparisons of score patterns or scores based on the same latent variables? Taking into account that method factors can have an impact on both the whole instrument and specific items of the instrument, four major types of bias can be distinguished. These are as follows: (1) cultural specificity of the theoretical variable, (2) domain underrepresentation, (3) method bias, and (4) item bias. Each of these four types of bias is discussed here in more detail (see also Table I).

Cultural Specificity of the Theoretical Variable

The first question asks what can cause bias theoretical variables. Here, the answer here lies in the cultural

specificity of the theoretical variable itself. When the theoretical variable and the framework in which it is embedded refer to phenomena that are culturally constructed, the use of that variable is limited to the culture concerned. This implies that there is no functional equivalence. For instance, "reading disabilities" make sense only in literate societies. This concept cannot be transported to illiterate societies. However, it is often not clear whether or not theoretical variables refer to culturally constructed aspects of psychological functioning, especially if they refer to the traits and processes that may be rooted in biological dispositions. Cultural relativists and universalists differ widely in the estimated *a priori* likelihood of culture-specific theoretical variables. According to the relativists, basic human traits and processes are fundamentally culturally constructed and show at best an echo of underlying biological processes. According to the universalists, these traits and processes are universal in human behavior, with the cultural context having an impact on the specific behavior manifestations in which they emerge.

Domain Underrepresentation

The second question asks about sources of bias in the use of the same instrument across cultural groups. Since the observed variables form the interplay between the stimuli of the measurement procedure and the behavior that is elicited by it, the causes might be situated (1) in the stimulus material or (2) in the fact that the intended behavior is not elicited by the stimulus material. Here, the first problem is examined. The second problem is discussed as method bias.

As has been made clear in the presentation of the general framework, the stimuli must be relevant to and representative of the domain to which they are supposed to refer. Since cultural groups can differ widely in their behavioral repertoire, this can pose serious problems. A set of stimuli that is relevant to and representative of the target domain in one cultural group need not be relevant and representative in another cultural group. An instrument constructed in one cultural group might contain items that are irrelevant for the corresponding domain in another group. For instance, an item about systematically locking one's windows and doors might be a good indicator of psychoticism in cold countries but not in warm countries, where windows must be opened at night to let in the fresh air. In addition, it is possible that the stimuli of an instrument designed in one cultural group are relevant, but not representative of the target domain in another cultural group. This is called domain underrepresentation. It means that the results of the measurement cannot be generalized to the whole domain. This implies that the same observed variables are insufficient or cannot be used across cultural groups to measure the same theoretical variable. For instance, in 1996, Ho demonstrated that the domain of behaviors relevant to the theoretical variable of filial piety is much broader in China than in the United States. Merely translating a U.S. instrument for filial piety seriously underrepresents the domain in China.

When the behavioral repertoire turns out to be very different between cultural groups, the question arises as to whether this points to noncomparability of the underlying theoretical variable. For instance, the question has arisen as to whether or not filial piety has a different meaning in a Chinese context than in a U.S. context and thus is not comparable between the two cultural groups. Hence, cultural specificity of the theoretical variable and domain underrepresentation are often discussed together under the umbrella term of construct bias.

Method Bias

The third and last question asks which factors can cause bias in quantitative comparisons of scores or score patterns between cultural groups. If cultural specificity of the theoretical variable and domain underrepresentation can be excluded, then the remaining type of bias is method bias. The possible threats to full score and metric equivalence are discussed together, because the factors that have been described in the literature to cause bias in the full comparability of scores (causing a lack of full score equivalence) also affect comparability of score patterns (causing a lack of metric equivalence). Moreover, the factors can also affect structural equivalence (see also Table I). Method bias refers to all those biasing effects that are caused by the specific method and context of the measurement. It must be noted that method bias does not affect cross-cultural comparisons, if it operates in the same direction and to the same extent within each cultural group. From a cross-cultural perspective, the problem lies in a differential impact of method bias across cultural groups.

In the literature, method bias is restricted to those factors that have a biasing effect on all or substantial parts of a measurement instrument. Method factors that have an effect only on specific items are treated separately as item bias. Although, conceptually, item bias is just a form of method bias, there is a good reason to distinguish the two. As seen in the next section, item bias can often be straightforwardly detected by applying adequate psychometric methods, whereas the detection of method bias requires additional research using different methods. According to a 1997 paper by Van de Vijver and Tanzer, factors that cause method bias relate to the stimulus material, how it is administered, and to whom it is administered. These are called, respectively, instrument bias, administration bias, and sample bias.

Instrument Bias

Instrument bias is caused by specific item content, specific response format, and specific response styles. Differential familiarity across cultural groups with either the item content or the response format can be considered a major source of method bias. Lack of familiarity with the stimulus material or the response format may cause unintended difficulties, whereas familiarity can lead to testwiseness and unintended test-easiness. For instance, it has been demonstrated that the direction of writing the alphabet (from the left to the right for Latin languages or from the right to the left for Arab languages) has an impact on the difficulty of items in a figural inductive reasoning test that are presented in a horizontal way. The differential impact of the response format was demonstrated in a 1979 study by Serpell. British children outperformed Zambian children in a pattern reproduction task with a paper-and-pencil response format, but not when plasticine or configurations of hand positions were used to present the items. The cultural difference in performance was even reversed when respondents had to reproduce patterns in iron wire; making toys in iron wire is a favorite pastime in Zambia.

Another form of instrument bias is caused by differences in response style. Specific response formats may have a differential impact on response styles, such as social desirability, extremity scoring, and acquiescence. For instance, Hispanics tend to use more the extremes of the scale than Anglo-Americans, but only when a 5-point scale is used, and not when a 10-point scale is used.

Administration Bias

Administration bias refers to all biasing effects that are caused by the way in which a test is administered. This bias may be due to a lack of standardization across cultural groups or by different interpretations of standardized administration procedures. A lack of standardization in test administration between cultural groups may be caused by (1) differences in physical conditions (such as temperature or luminosity) and social environment (such as class size when subjects are tested in groups); (2) different instructions for the respondents due to ambiguous guidelines or differential expertise of the test administrators; and (3) problems in the communication between tester and testee due to differences in language, the use of an interpreter, or culture-specific interpretation of the instructions.

Even if the test administration is perfectly standardized from an objective point of view, a differential meaning of the characteristics of the test administration may lead to bias. Effects have been reported in the literature of (1) the use of measurement or recording devices that arouse more curiosity or fear in cultural groups less acquainted with them and (2) differential tester (or interviewer or observer) effects. In particular, when the tester is of a different ethnic background as the testee, all kinds of unintended social processes can be elicited, such as defensive or more socially desirable responding.

Sample Bias

Sample bias is due to the noncomparability of cultural samples on other characteristics than the target variable. Sample bias threatens comparability, especially if it interacts with other forms of method bias. For instance, if cultural groups differ systematically in the motivation to respond to cognitive tests, the interpretation of differences in mean scores is ambiguous. These could point to differences in motivation, ability, or both. The differential impact of the nuisance factors across groups leads to a lack of full-score equivalence. However, even if the measurement instrument generates fully score-equivalent scores, sample bias can have an adverse impact on the interpretation of the results. For instance, if the cultural samples strongly differ in terms of their average educational level, observed differences on cognitive tests could be interpreted in terms of such educational factors, rather than in terms of cultural differences in cognitive functioning.

Item Bias

Item bias refers to those causes of noncomparability that are due to responses on specific items. The most obvious reason for item bias lies in a poor translation of the item. For instance, in an international study, the original English item used the term "webbed" feet for water birds. In the Swedish version, this item was translated as "swimming" feet, which caused an unintended easiness of this item for Swedish pupils. The impact of a poor translation may increase when the original item has an ambiguous meaning. Another problem is that items can cause nuisance variance by invoking unintended traits or processes. For instance, the word "dozen" in a numerical ability item might introduce additional verbal abilities. In addition, individual items might be more appropriate for the domain of interest in one cultural group than in another.

Psychometrically, the item bias can affect the intercept, the weight, or both parameters in the relationship between the observed and latent variable. Item bias that affects only the intercept is called uniform item bias. Within a particular cultural group, the item score (or log odds of a correct versus an incorrect answer) is systematically higher than in another cultural group independently of (or uniform across) the position on the latent variable; the bias is the same for all persons. Item bias that affects the weight parameter is called nonuniform item bias. If the weight parameter differs, the size of the bias for a respondent in a group depends on her or his position on the latent variable. The bias is thus not uniform across the possible positions on the latent variable.

Bias and the Level of Inference

If item bias is limited to a few items in a test, it can be detected in a fairly straightforward manner by means of psychometric modeling. Expected consistencies in item responses specified by the psychometric model will not hold for biased items across cultural groups. However, no evidence of bias will be detected if all items are uniformly biased in the same direction for a particular cultural group. Such a generalized uniform bias will be modeled psychometrically by an average difference in position on the latent variable between the cultural groups, rather than by a similar uniform bias in each of the items. It depends on the intended level of inference whether generalized uniform bias forms an alternative explanation for observed differences between cultural groups.

At a low level of inference, a measurement is basically focused on behavioral manifestations. This means that inferences stay close to the observed behavior. If item bias can be excluded by psychometric modeling, then differences between cultural groups can be interpreted in a straightforward way. Suppose one is interested only in whether children can solve correctly the type of items used in intelligence tests. After item bias has been ruled out psychometrically, a lower score in one cultural group means that the children of that group know less well how to solve the items.

At a higher level of inference, the measurement is focused on the traits and processes that underlie the concrete behavior, rather than on the behavior as such. Items are selected because they are assumed to form privileged phenomena where the underlying traits and processes manifest themselves clearly. At this level of inference, uniform bias forms a plausible alternative hypothesis for observed differences between cultural groups, even if psychometrically no item bias can be detected. Cultural groups can differ widely in their repertoire of behavior and thus also in the extent to which underlying processes and traits underpin concrete behavior in each of the groups. For instance, observed cultural and ethnic differences in scores on intelligence tests strongly relate to group differences in schooling and socioeconomic status and thus to group differences in behavioral repertoire. Factors relating to the differences in repertoire, such as stimulus unfamiliarity, may well have a generalized effect across items in an intelligence test. This is illustrated by the decrease during the past century of the gap in intelligence scores between Afrikaans- and English-speaking white South Africans as these two groups were growing closer together in schooling and socioeconomic status.

It can be concluded that higher level inferences are more susceptible to generalized bias factors than low-level inferences that focus on specific behavioral manifestations. Paradoxically, the absence of any form of item bias is usually easiest to establish if there are no observed differences at all between cultural groups.

Research and Data-Analytic Methods for Detecting Bias and Justifying Equivalence

This section presents some major methods that can be used to detect bias or justify an assumption of equivalence in cross-cultural measurement. Going back to the general framework, the question is how can one justify the use of the same theoretical, observed, and latent variables across cultural groups or how one can detect bias. Four major approaches follow from the framework on equivalence and bias, namely, (1) investigating the nomological network in order to justify functional equivalence or to detect cultural specificity of the theoretical variable; (2) focusing on the domain in order to justify relevance and representativeness of the items or to detect irrelevance and underrepresentation; (3) applying psychometric methods in order to justify structural, metric, and possibly full score equivalence or to detect differences in weight and intercept parameters; and (4) applying a multimethod approach in order to further justify structural, metric, or full score equivalence or to detect method bias (see also Table I).

Nomological Network

As already discussed, the network of convergent and discriminant relationships of a theoretical variable with other theoretical variables is one of the main sources of information on the scientific meaning of a theoretical variable. Therefore, the empirical study of the nomological network within each cultural group forms one of the important strategies to exclude cultural specificity of the theoretical variable and support functional equivalence.

The study of the nomological network, however, is interesting not only for justifying the identity of theoretical variables cross-culturally. It can also contribute to elucidating the meaning of cross-cultural differences obtained with full score equivalent measurements. In 1987, Poortinga and Van de Vijver compared the study of the impact of culture on behavior with the peeling of an onion. The variables from the nomological network that account for the cultural differences are like the layers of an onion. The impact of culture has been fully grasped when all cultural differences have disappeared after the effects of those variables are controlled for. For instance, in 1989, Earley found that a difference in social loafing (working less hard in a group than alone) between a U.S. sample and a Chinese sample disappeared when the allocentric versus idiocentric orientation of the individual

respondents participating in the study was controlled for. Thus, cultural differences in social loafing could be attributed to differences in allocentric versus idiocentric orientations.

Even if there are no clear *a priori* hypotheses about the explanation of observed cross-cultural differences, inclusion of variables that possibly belong to the nomological network, such as social, political, or economic context variables, can considerably strengthen cross-cultural research. Consistent relationships with these variables make an interpretation of cross-cultural differences in terms of mere method artifacts less likely to occur. Moreover, they can offer a rich source for generating new hypotheses about the meaning of the cross-cultural differences.

Domain

In the context of achievement testing, judgment methods are frequently used to detect irrelevance of items for specific ethnic or cultural groups. Expert judges who are familiar with the ethnic or cultural group screen items for inappropriate content. These items can then be deleted from the instrument.

In other research fields, relevance and representativeness of the stimuli can also be studied by means of key informants who are well acquainted with the local culture and language. For instance, in 1992, Schwartz asked local researchers to add culture-specific value items that they thought were not represented in the original instrument. Later analyses indicated that these culture-specific value items did not lead to culture-specific value dimensions, which supported the comprehensiveness of the original model.

Another approach is to study the domain of investigation in a rather open and unstructured way, such as performing content analysis on responses to open questions. For instance, in a 2002 comparative study of the cognitive structure of emotions between individuals from Indonesia and from The Netherlands, Fontaine and co-workers first asked subjects to write down as many emotion terms as they could think of in 10 minutes. Thus, they ensured that the emotion terms used were relevant to and representative of the domain of emotion terms in each of these groups.

Internal Structure of the Instrument

If only a few items in an instrument are biased, psychometric modeling of the internal structure of the item responses offers a powerful tool for identifying bias. Here, six prototypical data-analytic and psychometric models that allow for detection of item bias are presented, namely, (1) two-factorial analysis of variance (ANOVA), (2) exploratory factor analysis, (3) confirmatory factor analysis, (4) the Mantel-Haenszel statistic, (5) logistic regression, and (6) the two-parameter logistic model. Which of these models can be applied in order to detect uniform or nonuniform bias or to justify structural, metric, or full-score equivalence is also briefly mentioned.

These models can be classified according to two criteria. The first criterion is whether or not observed and latent variables are related to one another by means of a formal measurement model. The two-parameter logistic model, exploratory factor analysis, and confirmatory factor analysis all rely on a formal measurement model. If no formal measurement model is used, the position on the latent variable is estimated by a so-called proxy. ANOVA, the Mantel-Haenszel statistic, and logistic regression all rely on the sum of the individual item scores as a proxy for the position on the latent variable. The second criterion is the assumed measurement level of the observed and the latent variables (nominal, ordinal, interval, or ratio level). ANOVA, exploratory factor analysis, and confirmatory factor analysis can be used when both the observed and the latent variables are assumed to be measured at the interval level. The Mantel-Haenszel statistic, logistic regression, and the two-parameter logistic model can be used when the observed variables are measured at the nominal (dichotomous, correct–incorrect) level and the latent variable is assumed to be at the interval level.

ANOVA

If both the observed and the latent variables are supposed to be measured at the interval level, and the sum score is used as a proxy for the position on the latent variable, ANOVA with both the proxy and the cultural group as independent variables can be used to detect uniform and nonuniform item bias. First, the sum score is reduced to a limited number of score levels. Then, an ANOVA is performed for each item separately with score level and group as independent variables. The main effect for the score level gives information about the strength of the relationship between proxy and observed variable across cultural groups. If only this effect is significant, the item demonstrates full score equivalence. If the main effect for cultural group is significant, there is uniform bias; independent of the score level, the item has a higher mean in one cultural group than in the other cultural group. If only the main effect for score level and the main effect for cultural group are significant, the item still demonstrates metric equivalence. If the interaction effect between score level and cultural group is significant, then there is nonuniform bias.

Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a classical formal measurement model that is used when both observed and latent variables are assumed to be measured at the interval level. Characteristic of EFA is that the observed variables are first standardized (mean of zero and standard

deviation of 1). EFA is executed on the correlation matrix between the items. In EFA, a latent variable is called a factor and the associations between latent and observed variables are called factor loadings. Factor loadings are standardized regression weights. Since EFA is an exploratory technique, there is no expected distribution of loadings; hence, it is not possible to test statistically whether or not factor loadings are the same across cultural groups. However, congruence measures, such as Tucker's ϕ, have been developed to indicate whether the pattern of factor loadings across items on a factor is the same across cultural groups. Sufficient congruence for structural equivalence is usually taken to be found if Tucker's ϕ exceeds 0.95. Values below 0.90 are taken to indicate that one or more items show deviant factor loadings and thus show bias. Bootstrap procedures have been developed to test the identity of factor loadings in EFA.

EFA is used to investigate structural equivalence. However, since it works on standardized variables (mean of zero and standard deviation of 1), this model is not suited to detect nonuniform and especially uniform item bias.

EFA is often used in the multidimensional situation where more than one latent variable is measured at the same time. Before evaluating congruence in this case, the factor structures should be rotated toward a target structure.

Confirmatory Factor Analysis

CFA offers a measurement model based on structural equation modeling. It is related to EFA (latent variables are called factors and item weights are factor loadings), but does not suffer from several of the limitations of EFA for bias research. It is executed on the means and variance–covariance matrix instead of on the correlation matrix. It can thus detect both nonuniform and uniform bias. Moreover, it is an inferential model that allows for statistical testing of the model parameters. With CFA, structural, metric, and full-score equivalence can be modeled elegantly. Structural equivalence holds if the same factor model applies in each of the cultural groups. This means that each of the items has a significant loading on the predicted factor. Metric equivalence holds if the factor loadings are the same across cultural groups per item. Full score equivalence holds when both the factor loadings and the intercepts are the same per item. Like EFA, CFA is often used when there is an array of variables measuring more than one dimension.

Mantel-Haenszel Statistic

The Mantel-Haenszel statistic (MH) can be used for comparing two cultural groups when the observed item scores are dichotomous (correct–incorrect) and the sum score is used as a proxy for the latent variable. In a first step, the sum score is reduced to a limited number of score levels. Within each score level, the odds of correctly versus incorrectly responding are computed within each of the two groups and the ratio of these odds is taken. Then, the odds ratios are averaged across score levels. If there is no uniform bias present, the average odds ratio should equal 1. This means that the probability of correctly responding is on average the same in each group for subjects with the same position on the proxy. If the odds ratio deviates significantly from 1, there is a case for uniform bias. It means that in one cultural group the probability of correctly answering the item is systematically higher than in the other cultural group with the same position on the proxy. Because the MH is robust and easy to use, it is one of the most frequently used psychometric methods in item bias research.

The MH is useful only to demonstrate full score equivalence if structural equivalence and metric equivalence have already been demonstrated by means of other psychometric analyses. The MH as such gives no information about the relationship between item score and proxy within each group. Moreover, the MH is not powerful in detecting nonuniform bias, because it is based on the average odds ratio across score levels.

Logistic Regression

When the observed variables are dichotomous (correct–incorrect) and the sum score is used as a proxy for the latent variable, logistic regression can be used to detect both uniform and nonuniform bias. Per item a logistic regression analysis is executed. In this regression, the sum score across items is the independent variable and the logarithm of the odds of correctly versus incorrectly responding on a single item is the dependent variable. Equal intercepts and equal-weight parameters across cultural groups indicate full score equivalence of the item. Unequal intercepts across cultural groups point to uniform bias. If the weight parameters are the same across cultural groups, the item is characterized by metric equivalence. Unequal weights across cultural groups point to nonuniform bias.

Two-Parameter Logistic Model

The two-parameter logistic model has already been presented. It models the probability of success on a specific item as a logistic function of the position on the latent variable. For each item, there is an item-characteristic curve defined by two parameters, namely, the item difficulty (y_i), which corresponds to the position on the latent trait where the probability of success is 50%, and the item discrimination (x_i), which represents the steepness of the logistic curve. If both the difficulty and the discrimination parameters are the same across cultural groups, then conditions for full score equivalence have been met. If the item difficulty parameter is not equal across cultural groups, there is uniform item bias. If only the item

difficulty parameters are different across cultural groups, but not the item discrimination parameters, there is a case for metric equivalence. If the item discrimination parameters are significantly different across cultural groups, then there is nonuniform item bias. However, if the two-parameter logistic model holds within each of the cultural groups separately, then there is still a case for structural equivalence.

Multimethod Approach

In addition to the problem of generalized uniform bias for higher level inferences, which was discussed in the previous section, there is the difficulty that method factors can have a biasing effect on all item responses in a particular cultural group. For instance, the same items might be more susceptible to social desirability in one cultural group than in another. These systematic method effects may go unnoticed if only the internal structure of the instrument is analyzed. However, this type of bias can be detected by applying multiple methods or, more preferably, a multitrait–multimethod design. Only by applying different methods for measuring the same theoretical variable can systematic effects associated with a specific method be revealed.

Conclusions

Cross-cultural measurements can be distorted in many different ways and this leads to noncomparability of scores. The tenet of the entire bias and equivalence literature is that one cannot simply assume full score equivalence in cross-cultural measurement and interpret cross-cultural differences at face value. In reaction to the recognition of the plethora of possible biasing effects, extensive psychometric, methodological, and theoretical tools have been developed to deal with these effects. This arsenal of tools offers a range of possibilities to empirically justify the intended level of equivalence and draw valid conclusions from cross-cultural measurements.

See Also the Following Articles

Item and Test Bias • Measurement Error, Issues and Solutions

Further Reading

Berry, J. W., Poortinga, Y. H., Segall, M. H., and Dasen, P. R. (2002). *Cross-Cultural Psychology: Research and Applications*, 2nd Ed. Cambridge University Press, Cambridge, UK.

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

Cole, N. S., and Moss, P. A. (1989). Bias in test use. In *Educational Measurement* (R. L. Linn, ed.), 3rd Ed., pp. 201–219. Macmillan, New York.

Harkness, J. A., Van de Vijver, F. J. R., and Mohler, P. P. (2003). *Cross-Cultural Survey Methods.* Wiley, Hoboken, NJ.

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

Messick, S. (1989). Validity. In *Educational Measurement* (R. L. Linn, ed.), 3rd Ed., pp. 13–103. Macmillan, New York.

Millsap, R. E., and Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Appl. Psychol. Measur.* **17,** 297–334.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *Int. J. Psychol.* **24,** 737–756.

Reynolds, C. R. (1995). Test bias and the assessment of intelligence and personality. In *International Handbook of Personality and Intelligence* (D. H. Saklofske and M. Zeider, eds.), pp. 543–573. Plenum, New York.

Steenkamp, J.-B.E.M., and Baumbartner, H. (1998). Assessing measurement invariance in cross-national context. *J. Consum. Res.* **25,** 78–90.

Tanzer, N. K. (1995). Testing across cultures: Theoretical issues and empirical results [Special Issue]. *Eur. J. Psychol. Assess.* **11**(2).

Van de Vijver, F. J. R., (ed.) (1997) Cross-cultural psychological assessment [Special Issue]. *Eur. Rev. Appl. Phychol*. **47**(4).

Van de Vijver, F. J. R., and Leung, K. (1997). Methods and data analysis of comparative research. In *Handbook of Cross-Cultural Psychology, Vol. 1, Theory and Method* (J. W. Berry, Y. H. Poortinga, and J. Panday, eds.), 2nd Ed., pp. 257–300. Allyn and Bacon, Boston, MA.

Van de Vijver, F. J. R., and Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research.* Sage, Thousand Oaks, CA.

Van de Vijver, F. J. R., and Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *Eur. Rev. Appl. Psychol.* **47,** 263–279.