Cross-Cultural Research Methods in Psychology

Fons van de Vijver, Tilburg University, Tilburg, The Netherlands

This article is reproduced from the previous edition, volume 5, pp. 2999–3003, © 2011, Elsevier Ltd., with revisions made by the Editor.

Abstract

Cross-cultural studies involve persons from different countries and/or ethnic groups. One of the central methodological problems of these studies is bias, the generic term for multiple explanations of cross-cultural differences. Three different types of bias are distinguished, depending on whether the source of interpretation problems derives from the construct, method of the study, or specific items (called construct, method, and item bias or differential item functioning, respectively). Equivalence refers to the implications of bias on score comparability. Linguistic, structural, measurement unit, and full score equivalence are described. Issues in test translation (translation – back-translation, committee designs, decentering) are discussed. Common subject- and culture-sampling schemes in cross-cultural research are mentioned. The article ends with a discussion of issues in combining individual- and country-level characteristics.

Cross-cultural studies involve persons from different countries and/or ethnic groups; a defining characteristic is their comparative nature. Most studies employ quantitative methods of data collection and analysis. Studies of cultural topics that are noncomparative and apply a qualitative methodology can be found in sociology ('cultural studies,' e.g., Barker, 2003), cultural psychology (Greenfield et al., 2003), and cultural anthropology (Angrosino, 2006).

The range of instruments used in comparative studies is very broad, ranging from highly standardized psychological tests, to observation schedules, and free interviews. In many studies existing Western instruments (mental tests, survey questionnaires, personality inventories) are administered either in a new cultural context, or not adapted to enhance their cultural appropriateness (Hambleton et al., 2004).

If two persons from different cultural groups show different scores on a reliable and valid measure of subjective well-being, these score differences may refer to individual differences in subjective well-being. However, the score differences may also arise from differential social desirability or some other response style, inappropriate translation, or inadequacy of the item to measure well-being in both groups. The example illustrates a central problem in cross-cultural research: observed score differences are often susceptible to multiple explanations (Creswell, 2013). When the same instrument has been administered to persons from different ethnic groups, it cannot be taken for granted that the same scores obtained in different cultural groups have the same psychological meaning.

The ambiguity of interpretation is a consequence of the methodological nature of culture as an independent variable. In laboratory studies researchers randomly assign subjects to experimental treatments. The random assignment leads to a firm control of ambient variables; ideally an experimental and control group are matched on all outcome-relevant characteristics (e.g., personality characteristics and socioeconomic status), except for the treatment variable studied (*see* Internal Validity). However, like gender and other intrinsic subject characteristics, culture is not an experimental treatment that can be manipulated. Groups with a different cultural background tend to differ on a variety of outcome-relevant characteristics. These differences may constitute rival explanations of observed cross-cultural differences. Without precautions to rule out these rival explanations, observed cross-cultural differences are open to multiple interpretations. Findings in cross-cultural research are more convincing when rival explanations have been more adequately dealt with.

Bias is the generic name of an important family of rival explanations (*see* Psychometrics). It refers to the common problem in the assessment of nonequivalent groups that scores obtained in different cultural groups are not an adequate reflection of the groups' standing on the construct underlying the instrument. If scores are biased, their psychological meaning is group dependent and group differences in assessment outcome are to be accounted for, at least to some extent, by auxiliary psychological constructs or measurement artifacts. A closely related concept is equivalence which refers to the absence of bias and hence, to similarity of meaning across groups. The two concepts have somewhat different historical roots and areas of application. Whereas bias usually refers to nuisance factors, equivalence has become the generic term for metrical implications of bias.

Bias and equivalence are not inherent properties of an instrument but arise in a group comparison with a particular instrument. Score comparisons of groups that differ in more test-relevant aspects will show a higher susceptibility to bias.

Sources of Bias

There are three bias sources in cross-cultural research. The first is called construct bias; it occurs when the construct measured is not identical across groups. Work on filial piety (psychological characteristics associated with being a good son or daughter) provides a good example (e.g., Yeh and Bredford, 2003). The Western conceptualization is narrower than the Chinese, according to which children are supposed to assume the role of caretaker of their parents when these grow old and become needy. Construct bias precludes the cross-cultural measurement of a construct with the same measure. An

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23148-6

inventory of filial piety based on the Chinese conceptualization will cover aspects unrelated to the concept among Western subjects, while a Western-based inventory will leave an important Chinese aspect uncovered.

An important source of bias, called method bias, can result from sample incomparability, instrument characteristics, tester and interviewer effects, and the method (mode) of administration. Examples are differential stimulus familiarity in mental testing and differential social desirability in personality and survey research. Some sources of method bias can be dealt with by careful preparation of the assessment instrument and its instruction manual (e.g., proper test instruction with a clear specification of what is asked from participants, standardization of administration, and adequate training of testers and interviewers). Yet, it may be impossible to eliminate all outcome-relevant sample characteristics, in particular when the cultural distance of the countries involved is large. There are indications that a country's Gross National Product (per capita) is positively related to its mean score on mental tests and negatively to its mean score on social desirability. Particularly in comparisons of culturally highly dissimilar groups it may be hard or even impossible to eliminate the impact of sources of method bias such as sources familiarity and social desirability.

Finally, bias can be due to anomalies at item level (e.g., poor translations); this is called item bias or differential item functioning. According to a definition that is used widely in psychology, an item is biased if persons with the same standing on the underlying construct (e.g., they are equally intelligent) but coming from different cultural groups, do not have the same average score on the item. The score on the construct usually is derived from the total test score. If a geography test administered to pupils in Poland and Japan, contains the item 'What is the capital of Poland?' Polish pupils can be expected to show higher scores on the item than Japanese students, even when pupils with the same total test score would be compared. The item is biased as it favors one cultural group across all test score levels. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Obinne and Amali, 2014).

An overview of common ways of addressing bias is given in Table 1.

Types of Equivalence

Elaborating on categorizations in the literature, four different types of equivalence are proposed here. The first type is labeled construct inequivalence. It amounts to comparing apples and oranges (e.g., the comparison of Chinese and Western filial piety, discussed above). Comparisons lack an attribute for comparison, also called tertium comparationis (the third term in the comparison). The second is called structural (or functional) equivalence. An instrument, administered in different cultural groups, shows structural equivalence if it measures the same construct in both groups (e.g., Raven's Standard Progressive Matrices test has been found to measure intelligence in various cultural groups). Exploratory factor analyses followed by target rotations or confirmatory factor analysis of correlation matrices (structural equation modeling) may be applied to examine structural equivalence. Structural equivalence does not presuppose identity of measures across groups. The measures may use different stimuli across groups. If different operationalizations have been chosen, structural equivalence can be examined by comparing nomological networks across groups.

The third type of equivalence is called measurement unit equivalence. Instruments show this type of equivalence if their measurement scales have the same units of measurement and a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same measurement units in each culture). It applies when a bias factor with a fairly

Table 1 Strategies for identifying and dealing with bias in cross-cultural research

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 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 (e.g., social desirability, acquiescence)
	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

Source: Van de Vijver and Tanzer, 1997. European Review of Psychological Assessment (reproduced with the permission of Swets and Zeitlinger).

uniform influence on the items of an instrument affects test scores of different cultural groups in a differential way. Social desirability and stimulus familiarity may exert this influence. Observed group differences in scores are then a mixture of valid cross-cultural differences and measurement artifacts. When the relative contribution of both sources cannot be estimated, the interpretation of group comparisons of mean scores remains ambiguous. Multigroup comparisons of confirmatory factor analytic models have been utilized to examine measurement unit equivalence. This technique, based on comparisons of covariance matrices, can identify bias sources that affect the covariance of items or tests but it cannot differentiate between valid differences in mean scores and bias sources with a uniform influence on all parts of an instrument.

Only in the case of scalar (or full-score) equivalence can direct comparisons be made; it is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different. This type of equivalence assumes the same interval or ratio scales across groups. Conclusions about which of the latter two types of equivalence applies are often difficult to draw and controversial. For example, racial differences in intelligence test scores have been interpreted as due to valid differences (scalar equivalence) and as reflecting measurement artifacts (measurement unit equivalence). Scalar equivalence assumes that the role of bias can be safely neglected. The demonstration of scalar equivalence draws on inductive argumentation. Therefore, it is easier to disprove than to prove scalar equivalence. This can be made plausible by measuring presumably relevant sources of bias (such as stimulus familiarity or social desirability) and showing that they cannot statistically explain observed cross-cultural differences in a multiple regression or covariance analysis.

The distinction between the latter two types of equivalence is immaterial when comparing experimental conditions or changes across cultures (e.g., developmental trajectories or training effects). In respect of scales that show measurement unit equivalence, measure changes at the level of full score equivalence.

Structural, measurement unit, and scalar equivalence are hierarchically ordered. The third presupposes the second, which presupposes the first. Moreover, higher levels of equivalence are more difficult to establish. It is easier to demonstrate that an instrument measures the same construct in different cultural groups (structural equivalence) than to demonstrate numerical comparability across cultures (scalar equivalence). On the other hand, higher levels of equivalence allow for more detailed comparisons of scores across cultures. Whereas in the case of structural equivalence, the only factor on which structures and nomological networks can be compared, scalar equivalence allows for more fine-grained analyses of cross-cultural similarities and differences, such as comparisons of mean scores across cultures in *t* tests and analyses of (co)variance.

Linguistic Equivalence

The concept of linguistic equivalence is developed in multilingual studies. Versions of an instrument in different languages show linguistic equivalence if these have the same characteristics that are relevant for the measurement outcome such as meaning, connotations of words and sentences, comprehensibility, and readability. Linguistic equivalence can be jeopardized by various sources, such as incorrect translations, words that are hard or impossible to translate (e.g., the English expression 'distress' does not have an equivalent in many languages), idiomatic expressions and metaphors (e.g., 'feeling blue'), and imprecise quantifiers ('rather often').

Because linguistic equivalence is not always guaranteed by a literal translation, it has become increasingly popular to utilize adaptations. In an adaptation, parts are changed (instead of literally translated) with the aim to improve an instrument's suitability for a target group.

Many studies employ a translation–back-translation procedure (Sperber, 2004). This amounts to a forward translation, followed by an independent back-translation and a comparison of the original and back-translated version, possibly followed by some alterations of the translation. Such a procedure provides a powerful tool to enhance the correspondence of original and translated versions that is independent of the researcher's knowledge of the target language. Yet, it also has some disadvantages. It puts a premium on literal reproduction; this may give rise to a stilted language use in the target version that lacks the readability and natural flow of the original. The problem may be compounded by translators' awareness of their involvement in a translation–back-translation procedure.

A second problem involves translatability. The use of idiom (e.g., the English 'feeling blue') or references to cultural specifics (e.g., references to country-specific public holidays) or other features that cannot be represented adequately in the target language challenges translation–back-translations designs (and indeed all procedures in which existing instruments are translated). During the 1990s there was a growing awareness that translations and adaptations require the combined expertise of social and behavioral scientists (with a competence in the construct studied) and experts in the target language(s) and culture(s). In this so-called committee approach in which the expertise of all relevant disciplines is combined, there is usually no formal accuracy check of the translation. Usage of the committee approach is popular in large international bodies in which texts are translated in many languages like the United Nations and the European Union.

When versions in all languages are developed simultaneously, a procedure called 'decentering' can be used. No single instrument or cultural group is then taken as starting point; individuals from different cultures jointly develop an instrument, thereby reducing the risk of introducing unwanted references to a specific culture.

The judgmental evidence of the designs to enhance linguistic equivalence can be easily combined with statistical approaches to establish equivalence. Reports of multilingual studies often provide a combination of judgmental–linguistic and empirical–statistical evidence to demonstrate the adequacy of an instrument and its translation. There is a tradeoff between the type of translation and the level of equivalence that can be obtained. When most or all questions of an instrument are adapted, structural equivalence is the highest level possible. When most or all items are literally translated, measurement unit and structural equivalence can be obtained. Recent statistical advancements in item response theory and structural equation modeling have made it possible to retain scalar equivalence,

even when not all stimuli are literally translated (provided that all items measure the same underlying construct in each group).

Sampling Cultures and Subjects

Cross-cultural studies can apply three types of schemes to sample cultures. Three types of sampling can be envisaged. The first is probability (or random) sampling. Because of the large cost of a probability sample from all existing cultures, it often amounts to stratified (random) sampling of specific cultures (e.g., Western cultures). The second and most frequently observed type of culture sampling is convenience sampling. The choice of cultures is governed here by availability and cost efficiency: researchers decide to form a research network and all participants collect data in their own country. In the third type, called systematic sampling, the choice of cultures is more based on substantive considerations. A culture is deliberately chosen because of some characteristic, such as in Nisbett and Miyamoto's (2005) study.

In survey research there is a well-developed theory of subject sampling (Scheaffer et al., 2011). In the area of cross-cultural research three types of sampling procedures of individuals are relevant as they represent different ways of dealing with confounding characteristics. The first is probability sampling. It consists of a random drawing from a list of eligible units such as persons or households. Confounding variables are not controlled for. The second type is stratified sampling. A population is stratified (e.g., in levels of schooling or socioeconomic status) and within each stratum a random sample is drawn. The purpose of stratification is the control of confounding variables (e.g., matching on number of years of schooling). The procedure cannot be taken to adequately correct for confounding variables when there is little or no overlap of the cultures (e.g., comparisons of literates and illiterates). The third procedure combines random or stratified sampling with the measurement of control variables. The procedure enables a statistical control of ambient variables (e.g., using an analysis of covariance).

Individual- and Country-Level Studies

The 1990s saw a growing interest in studies combining individual- and country-level data. Two kinds of studies have been reported. In multilevel hierarchical models (Raudenbush and Bryk, 2002), a regression model is used to explain individual variation (e.g., explaining pupils achievement scores on the basis of their intelligence) and class, district, or country variation (e.g., the explanation of the relationship between intelligence and achievement by means of school quality indicators). The second line of research involves the study of the same phenomena at different levels of aggregation (multilevel covariance structure analysis or multilevel factor analysis, Hox, 2002). From a conceptual point of view this line is more complicated, because it is well documented that (dis) aggregation can lead to methodological artifacts such as the ecological fallacy (*see* Ecological Fallacy, Statistics of), the incorrect application of culture-level characteristics to individuals. In each country a proportion of women are pregnant, but obviously, this proportion does not apply to any individual woman. In addition, equivalence issues have to be dealt with (Van de Vijver and Poortinga, 2002). Hofstede's (2001) famous study of values is based on country characteristics; the four dimensions he reported (individualism, power distance, uncertainty avoidance and masculinity–femininity) are characteristics of countries and their applicability to individual behavior cannot be taken for granted and has to be established. Multilevel covariance structure and factor analysis are powerful tools to examine the equivalence of phenomena at different levels of aggregation, answering questions such as the structural (in)equivalence of concepts like individualism–collectivism at individual and country level. For these and various other concepts the structural equivalence at different levels of aggregation is still unresolved.

See also: Psychometrics.

Bibliography

Angrosino, Michael V., 2006. Doing Cultural Anthropology: Projects for Ethnographic Data Collection. Waveland Press.

Barker, Chris, 2003. Cultural Studies: Theory and Practice. Sage.

Creswell, John W., 2013. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage.

Greenfield, Patricia M., et al., 2003. Cultural pathways through universal development. Annual review of psychology 54 (1), 461–490.

Hambleton, Ronald K., Merenda, Peter F., Spielberger, Charles D. (Eds.), 2004. Adapting Educational and Psychological Tests for Cross-cultural Assessment. Psychology Press.

Hofstede, Geert H., 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations. Sage.

Hox, Joop J., 2002. Multilevel Analysis: Techniques and Applications. Psychology Press.

Nisbett, Richard E., Miyamoto, Yuri, 2005. The influence of culture: holistic versus analytic perception. Trends in cognitive sciences 9 (10), 467–473.

Obinne, A.D.E., Amali, A.O., 2014. Differential item functioning: the implication for educational testing in Nigeria. International Review of Social Sciences & Humanities 7 (1).

Raudenbush, Stephen W., Bryk, Anthony S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, vol. 1. Sage.

Scheaffer, Richard, et al., 2011. Elementary Survey Sampling. Cengage Learning.

Sperber, Ami D., 2004. Translation and validation of study instruments for cross-cultural research. Gastroenterology 126, S124–S128.

Van de Vijver, Fons J.R., Poortinga, Ype H., 2002. Structural equivalence in multilevel research. Journal of Cross-cultural Psychology 33 (2), 141–156.

Yeh, Kuang-Hui, Bedford, Olwen, 2003. A test of the dual filial piety model. Asian Journal of Social Psychology 6 (3), 215–228.