Bias and equivalence in cross-cultural assessment

van de Vijver, F.J.R.; Tanzer, N.K.

Published in:
European Review of Applied Psychology = Revue Européenne de psychologie appliquée

Publication date:
1997

European Review of Applied Psychology, 4th trimester, vol. 47, n° 4, pp. 263 à 279

Bias and Equivalence in Cross-Cultural Assessment : An Overview

Fons van de Vijver[(1)], Norbert K. Tanzer[(2)]

[(1)]Tilburg University, Tilburg, the Netherlands
[(2)]University of Graz, Austria

SUMMARY

In every cross-cultural study, the question as to whether test scores obtained in different cultural populations can be interpreted in the same way across these populations has to be dealt with. Bias and equivalence have become the common terms to refer to the issue. A taxonomy of both bias and equivalence is presented. Bias can be engendered by the theoretical construct (construct bias), the method such as the form of test administration (method bias), and the item content (item bias). Equivalence refers to the measurement level at which scores can be compared across cultures. Three levels of equivalence are possible : the same construct is measured in each cultural group but the functional form of the relationship between scores obtained in various groups is unknown (structural equivalence), scores have the same measurement unit across populations but have different origins (measurement unit equivalence), and scores have the same measurement unit and origin in all populations (full scale equivalence). The most frequently encountered sources of bias and their remedies are described.

RÉSUMÉ

Dans toute étude interculturelle, il faut résoudre la question de savoir si les scores au test obtenus dans des populations culturellement différentes peuvent être interprétés de la même manière dans ces populations. Les termes de biais et d'équivalence sont ceux devenus habituels quand on envisage ce problème. On propose une taxonomie tant du biais que de l'équivalence. Le biais peut être produit par le construct théorique (biais de construct), par la méthode, par ex. par la forme d'administration du test (biais de méthode), et par le contenu d'item (biais d'item). L'équivalence se rapporte au niveau de mesure auquel les scores peuvent être comparés dans les cultures. Trois niveaux d'équivalence sont possibles : le même construct est mesuré dans chaque groupe culturel mais l'aspect fonctionnel de la relation entre les scores obtenus dans les différents groupes est inconnu (équivalence structurelle) ; les scores ont la même unité de mesure dans les populations mais ont différentes origines (équivalence d'unité de mesure) ; les scores ont la même unité de mesure et la même origine dans toutes les populations (équivalence d'échelle complète). Les sources de biais les plus fréquemment rencontrées sont décrites ainsi que les moyens d'y remédier.

Key words :

Bias,
equivalence,
construct bias,
method bias,
item bias,
overview.

Mots clés :

Biais,
équivalence,
biais de construct,
biais de méthode,
biais d'item,
revue.

This article will discuss bias and equivalence in cross-cultural assessment. We will start with a taxonomy of bias and equivalence (cf. Van de Vijver & Leung, 1997a, 1997b). A lot of cross-cultural research involves the application of instruments in various linguistic groups. Thus, the types of multilingual studies and their impact on bias and equivalence are discussed in the second section. The third section describes common sources of bias. The question of how to identify and to remove bias is discussed in the fourth section. Finally, conclusions are presented.

Bias and equivalence : definitions and taxonomy

Bias

Suppose that a geography test contains the item "What is the capital of Poland?" This test is administered to pupils in a large international educational achievement survey. The proportion of correct answers to the item will depend on, among other things, the pupils´ level of intellectual abilities, the quality of their geography education, and the distance of their country to Poland. Assuming that samples have been carefully composed, the question will enable an adequate comparison of the differences in knowledge of this particular item across all countries. However, suppose that the domain of the test is broader and that this item is used to assess geographical knowledge. Distance of the country to Poland will now become a nuisance variable. Pupils from central Europe are put at an advantage in comparison with pupils from, say, Australia and the U.S.A. Such problems, known as bias, are common in cross-cultural assessment. More generally, bias occurs if score differences on the indicators of a particular construct (e.g., percentage of students knowing that Warsaw is Poland's capital) do not correspond to

differences in the underlying trait or ability (e.g., geography knowledge). Inferences based on biased scores are invalid and often do not generalize to other instruments measuring the same underlying trait or ability. Equivalence can be defined as the opposite of bias. However, historically, they have slightly different roots and as a consequence, they have become and remained associated with different aspects of cross-cultural score comparisons. Bias has become the generic term for nuisance factors in cross-cultural score comparisons whereas equivalence tends to be more associated with measurement level issues in cross-cultural score comparisons. Both bias and equivalence are pivotal concepts in cross-cultural assessment. Equivalence of measures (or lack of bias) is a prerequisite for valid comparisons across cultural populations.

The above example may well serve to illustrate an important characteristic of bias and equivalence : Both concepts do not refer to intrinsic properties of an instrument but to characteristics of a cross-cultural comparison of that instrument. Statements about bias always refer to applications of an instrument in a particular cross-cultural comparison. An instrument that reveals bias in a comparison of German and Japanese individuals may not show bias in a comparison of German and Danish subjects.

The history of psychology has shown various examples of sweeping generalizations about differences in abilities and traits of cultural populations which, upon close scrutiny, were based on psychometrically poor measures. In order to avoid making such sweeping statements which may attract much initial attention but which eventually do a disservice to the field, the absence of bias (i.e., equivalence) should be demonstrated instead of simply assumed (Poortinga & Malpass, 1986).

In order to facilitate the examination of bias, the following taxonomy may be useful. Three kinds of bias are distinguished here (Van de Vijver & Leung, 1997a, 1997b ; Van de Vijver & Poortinga, 1997). The first one is construct bias. It occurs if the construct measured is not identical across cultural groups. Western intelligence tests provide a good example. In most general intelligence tests, there is an emphasis on reasoning, acquired knowledge, and memory. Social aspects of intelligence are often less emphasized. However, there is ample empirical evidence that the latter aspects may be more prominent in non-Western settings (e.g., Super, 1983). The term "intelligence" as commonly applied in psychology does not do justice to its specific domain of application which is education. Binet's assignment, to design a test to detect children with learning problems which led to the development of intelligent tests as we know them, is still discernible. The domain of the tests would be more appropriately called "scholastic intelligence."

A second example of construct bias can be found in the work on filial piety (i.e., behaviors associated with being a good son or daughter ; Ho, 1996). Compared to Western societies, children in Chinese societies have more and different obligations towards their parents. The difference may be caused by education and income. Kagitcibasi (1996) found in Turkey that "help with household chores" lost salience for parents with increased education. Similarly, the value of children as old age security for the parents decreases with the level of income. Therefore, a comparison of filial piety across cultural populations is susceptible to construct bias. When based on a Western conception, the instrument will not cover all relevant aspects in a non-Western context. Analogously, an instrument based on a Chinese concept will contain behaviors such as the readiness to take care of one's parents financially in their old age which are only marginally related to the Western concept of filial piety. When based on a collectivist notion, the instrument will be overinclusive and will contain various items that may well show little interpersonal variation and induce a poor reliability of the instrument in a Western culture.

The question to be asked is how to deal with construct bias : Is it possible to compare filial piety between individuals living in Western and non-Western cultures? Probably, the easiest solution is to specify the theoretical conceptualization underlying the measure. If the set of relevant Western behaviors is a subset of the non-Western set, then the comparison can be restricted to the Western set while acknowledging the incompleteness of the measure for the non-Western group.

The second type is method bias. The term "method bias" is coined because it derives from aspects described in the Method section of empirical papers. Three types of method bias can be envisaged. First, incomparability of samples on aspects other than the target variable can lead to method bias (sample bias). For instance, cultural groups often differ in educational background and, when dealing with mental tests, these differences can confound real population differences on a target variable. Intergroup differences in motivation can be another source of method bias caused by sample incomparability. For instance, subjects who have been frequently exposed to psychological tests will show less motivation than subjects for whom the instrument and/or the test situation has high novelty.

Method bias also refers to problems deriving from instrument characteristics (instrument bias). A well-known example is stimulus familiarity. Deregowski and Serpell (1971) asked Scottish and Zambian children in one condition to sort miniature models of animals and motor vehicles, and in another condition to sort photographs of these models. Although no cross-cultural differences were found for the actual models, the Scottish children obtained higher scores than the Zambian children when photographs were sorted. Response procedures can also induce method bias. Serpell (1979) asked Zambian and British children to reproduce a pattern using paper-and-pencil, plasticine, configurations of hand positions, and iron wire (making models with iron wire is a popular pastime among Zambian boys). The British scored significantly higher in the paper-and-pencil procedure while the Zambians scored higher when iron wires were utilized. An example using questionnaires can be found in the work by Hui and Triandis (1989). Hispanics tended to choose extremes on a five-point rating scale more often than White Americans. This tendency was, however, not found when a ten-point scale was used.

A final type of method bias arises from administration problems (administration bias). Communication problems

between interviewers and interviewees can easily occur, especially, when they have different first languages and cultural backgrounds (cf. Gass & Varonis, 1991). Interviewees´ insufficient knowledge of the testing language and inappropriate modes of address or cultural norm violations on the part of the interviewer (e.g., Goodwin & Lee, 1994) can seriously endanger the collection of appropriate data.

Method bias can have devastating consequences on the validity of cross-cultural comparisons. Method bias will often lead to a shift in average scores. For example, stimulus familiarity and social desirability tend to influence all items of an instrument and, hence, they will induce a change in average scores. Such a change may occur independently of possible cross-cultural differences on the target variable. For example, suppose that attitudes towards soft drug use are measured among youngsters in France and the Netherlands using face-to-face interviews. The answers given by the youngsters may well be influenced by the more restrictive laws surrounding soft drug use in France as compared to the Netherlands. The question about possible cross-national differences in attitudes towards drug use can be confounded by differential social desirability. In general, method bias can affect observed intergroup differences. Attempts to disentangle method bias and valid cross-cultural differences are anything but trivial. Neglecting the impact of method bias can seriously threaten the validity of inferences. In the example above, if the assessment method used is less prone to induce social desirability (e.g., an anonymous administration of the questionnaire in large groups), another pattern of French-Dutch differences may be observed.

A final type is item bias or differential item functioning (e.g., Berk, 1982 ; Holland & Wainer, 1993). Unlike construct and method bias, item bias refers to distortions at item level. Biased items have a different psychological meaning across cultures. Suppose that the subjects´ responses in one cultural group are partly determined by social desirability for one item of a self-report inventory. Then, a comparison of total test scores across cultures would be invalid when this item is included. Item bias has received considerable attention in the literature ; most studies of bias focused on exploring and testing statistical procedures to identify item bias. From a statistical-methodological perspective, an item is taken to be biased if persons from different groups with the same score on the construct, commonly operationalized as the total score on the instrument, do not have the same expected score on the item (Shepard, Camilli, & Averill, 1981). For persons from different cultural groups with equal total scores (i.e., persons from different cultural groups who are equally intelligent, anxious or whatever is measured), an unbiased item should be equally difficult (or attractive). Thus, they should have equal mean scores across the cultural groups ; different means on that item point to item bias.

Many statistical techniques are available to detect item bias (cf. Holland & Wainer, 1993 ; Van de Vijver & Leung, 1997a, 1997b). The most popular procedure is the Mantel-Haenszel statistic (cf. Holland & Thayer, 1988 ; Klieme & Stumpf, 1991). It is a procedure for analyzing bias in dichotomously scored items which are common in mental tests. Suppose that a test of 6 items had been administered to two cultural groups of 1000 persons each. In the first step of the Mantel-Haenszel procedure, both samples will be split up in subgroups with equal test scores. Subjects solving either none or all items do not provide information as to whether an item is biased and must, therefore, be excluded from the Mantel-Haenszel bias analyses. Thus, within each culture, the first subgroup will consist of subjects with a total test score of one, the next subgroup with a total score of two, and so on. The Mantel-Haenszel procedure then compares the averages of the items across score groups. An unbiased item will show averages that, for all score groups, are equal across cultures. Items that are easier or more difficult in most or all score groups of one of the cultures are taken to be biased.

Equivalence

It has become customary to treat equivalence from a measurement level perspective. We will also do this here and adopt the levels of equivalence proposed by Van de Vijver and Leung (1997a, 1997b ; see Poortinga, 1989, for a similar approach). A distinction can be made between hierarchically linked types of equivalence. The first is construct equivalence (also labeled structural equivalence and functional equivalence). It means that the same construct is measured across all cultural groups studied, regardless of whether or not the measurement of the construct is based on identical instruments across all cultures. It implies the universal (i.e., culture-independent) validity of the underlying psychological construct and, in a terminology frequently used in cross-cultural psychology (cf. Triandis & Marín, 1983), can be associated with an "etic" position. Construct inequivalence, on the other hand, will be observed when an instrument measures different constructs in two cultural groups (i.e., when "apples and oranges are compared") or when the concepts of the construct overlap only partially across cultures. It may also result when constructs are associated with different behaviors or characteristics across cultural groups ("cultural specifics"). The assumption of construct inequivalence can be associated with an "emic" position which emphazises the idiosyncrasies of each culture and, as a consequence, favors an indigenous approach to assessment.

As an example of construct equivalence, suppose that a researcher is interested in traits and behaviors associated with loneliness in Austria and China. The study could begin with a local survey in which randomly chosen adults are asked to generate such traits and behaviors. If the lists generated are essentially identical across cultures, a loneliness questionnaire with identical questions in the two countries could be composed. Data obtained with the instrument in the two countries can be subjected to exploratory or confirmatory factor analyses in order to examine construct equivalence (cf. Van de Vijver & Leung, 1997a, 1997b). If there are major differences in the traits and behaviors, one will need to tailor the measure to the cultural context. This means that at least some items will be different in the two countries. The construct equivalence of measures should

then be addressed in a more indirect way. A common procedure is to examine the nomological network of the measure (Cronbach & Meehl, 1955). Does the measure show a pattern of high correlations with related measures (convergent validity) and low correlations with measures of other constructs (discriminant validity) as would be expected from an instrument measuring loneliness?

The next level of equivalence is called measurement unit equivalence. This level of equivalence can be obtained when two metric measures have the same measurement unit but have different origins. In other words, the scale of one measure is shifted with a constant offset as compared to the other measure. An example can be found in the measurement of temperature using Kelvin and Celsius scales. The two scales have the same unit of measurement, but their origins differ 273 degrees. Scores obtained with the two scales cannot be directly compared but if the difference in origin (i.e., the offset) is known, their values can be converted so as to make them comparable. In the case of cross-cultural studies with measurement unit equivalence, no direct score comparisons can be made across cultural groups unless the size of the offset is known (which is rarely the case), but differences obtained within each group can still be compared across groups. For example, gender differences found in one culture can be compared with gender differences in another culture for scores showing measurement unit equivalence. Likewise, change scores in pretest-posttest designs can be compared across cultures for instruments with measurement unit equivalence.

The highest level of equivalence is scalar equivalence or full scale equivalence. This level of equivalence can be obtained when two metric measures have the same measurement unit and the same origin. For instance, when temperature is measured using a Celsius scale (which is of interval level) in both groups, differences in temperature can be compared directly between the two groups.

The distinction between measurement unit and scalar equivalence is important in cross-cultural research. The latter assumes completely bias-free measurement. Bias tends to challenge and can lower the level of equivalence. Construct bias leads to conceptual inequivalence. As a consequence, instruments that do not adequately cover the target construct in one of the cultural groups cannot be used for cross-cultural score comparisons. On the other hand, method and item bias will not affect construct equivalence. Construct equivalence implies only that the same construct is measured across cultures. If no direct score comparisons are intended across cultures, neither method nor item bias will be a threat to cross-cultural equivalence. However, method and item bias can seriously threaten scalar equivalence. An item systematically favoring a particular cultural group will obscure the underlying real cross-cultural differences in scores on the construct. Therefore, such a bias will reduce scalar equivalence to measurement unit equivalence.

The debate about cross-cultural differences in cognitive test performance can be largely seen as a debate about the level of equivalence of cross-cultural score comparisons. For example, Jensen (1980) argues that when appropriate instruments are used (he mentions the Raven test as an example), cross-cultural differences in test performance reflect valid intergroup differences and show full scalar equivalence. Mercer (1984), on the other hand, states that common intelligence tests show problems such as differential familiarity and that this method bias will only allow measurement unit equivalence. The obvious implication is that group differences in the Raven scores reflect differences in intellectual abilities according to Jensen's reasoning while group differences mainly or exclusively reflect method bias in Mercer's reasoning.

Multilingual studies : translation

It is widely accepted that the translation of psychological instruments involves more than rewriting the text in another language (Bracken & Barona, 1991 ; Brislin, 1980, 1986 ; Geisinger, 1994 ; Hambleton, 1994). An appropriate translation requires a balanced treatment of psychological, linguistic, and cultural considerations (Hambleton, 1994 ; Van de Vijver & Hambleton, 1996).

There are two common procedures to develop a translation. First, there is the translation-backtranslation procedure (Werner & Campbell, 1970). A text is translated from a source into a target language ; a second interpreter (or group of interpreters) independently translates the text back into the source language. The accuracy of the translation is evaluated by comparing the original and backtranslated versions. The procedure has been widely applied and it can identify various kinds of errors. However, a translation that is linguistically correct may still be of poor quality from a psychological point of view. A nice example given by Hambleton (1994, p. 235) is the test item "Where is a bird with webbed feet most likely to live?" The Swedish translation of the English "bird with webbed feed" into "bird with swimming feet" provides a much stronger clue to the solution than the English original item.

This problem, which would most likely remained undetected during a translation–back translation procedure, may be detected by the second procedure, the committee approach. A group of people, often with different areas of expertise (such as cultural, linguistic, and psychological) prepare a translation. The major strength of the committee approach is the cooperative effort that can improve the quality of translations and, especially, in the case when the committee members have complimentary areas of expertise.

The procedure chosen to obtain an adequate translation will depend on whether a new instrument is to be developed or whether an existing instrument is to be translated to be used in a multilingual context. The former is known as simultaneous development and the latter as the successive development of different language versions. From a methodological perspective, the first option is often easier to carry out because typical problems of successive development such as the use of local idioms which are difficult to translate, can often be easily avoided. Still, the most common practice in multilingual studies is to use successive development.

Three options are available to researchers in the successive development method (Van de Vijver & Leung,

1997a, 1997b). The first is application. It amounts to the literal translation of an instrument into a target language. In this option, it is implicitly assumed that the underlying construct is appropriate in each cultural group and that a simple, straightforward translation will suffice to get an instrument that adequately measures the same construct in the target group. The literal translation is by far the most common option in test translations.

The second option is adaptation. For some instruments, it is unrealistic to assume that a simple translation will yield an instrument that will adequately cover the same construct in the target group. For example, a measure of anxiety may contain some items that can well be translated but may require the rewording of other items to ensure that culturally idiosyncratic expressions of the construct are included. An adaptation amounts to the literal translation of a part of the items and/or changes in other items and/or the creation of new items. Adaptations are based on the notion that the use of the application option would yield biased instruments. For example, a core of common items may show construct bias because they poorly sample the domain of possible items in at least one culture and, hence, the construct is insufficiently represented. A good example for applying the adaptation option is the State-Trait Anxiety Inventory (STAI ; Spielberger, Gorsuch, & Lushene, 1970). This instrument had been adapted into more than 40 languages. The various language 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, were measured adequately in each language (e.g., Laux, Glanzmann, Schaffner, & Spielberger, 1981).

Finally, in some cases, the instrument has to be adaptated to such a degree that practically a new instrument is assembled. Hence, this third option is called assembly. In particular, when construct bias caused by differential appropriateness of the item content for the majority of the items threatens a direct comparison, assembly may be an adequate option. Another indication for using the assembly option would be the incomplete overlap of the construct definition across cultures (e.g., aspects of the construct that are salient for some cultures but are not covered in the instrument). According to Church (1987), Western personality instruments do not cover indigenous personality constructs of the Filipino culture. He formulated directions for the construction of a culturally more appropriate personality instrument. Cheung, Leung, Fan, Song, Zhang, and Chang (1996) argued that adapting a Western personality measure would not capture all the relevant dimensions of personality in the Chinese culture. They developed the Chinese Personality Assessment Inventory which contains several indigenous personality dimensions such as "face" and "harmony."

It may be clear from the description that the three translation options differ in the amount of items that can be retained in the translation process. Going from the first to the third option, an increasing number of items will be changed in the translation process.

The choice of the translation option has implications for the expected level of equivalence. Assembly will preclude numerical score comparisons across cultures because they require scalar equivalence and construct equivalence is the highest level of equivalence possible. In many cases this is exactly the main question of the translation : Do the results obtained with regard to the nomological network show that the same psychological construct has been measured in each cultural group?

From a statistical perspective, adaptations are the most cumbersome. Direct score comparisons will be forbidden because these are not based on the same instrument. One could restrict the score comparisons to the items common in all cultural groups. In general, this will constitute an unsatisfactory solution because the rest of the items will not be used. Moreover, when the set of common items is small, they will not adequately cover the construct, and score comparisons will suffer from a low ecological validity and a poor generalizability to more appropriate measures of the construct. Fortunately, there are statistical techniques such as item response theory (e.g., Hambleton & Swaminathan, 1985 ; Hambleton, Swaminathan, & Rogers, 1991) that allow score comparisons of persons' abilities or traits even when the items of an instrument are not entirely identical. When these techniques are applied, the possibility of scalar equivalence is maintained. If one wants to examine construct equivalence, the use of structural equation models may be considered (cf. Byrne, 1989, 1994). Confirmatory factor analysis allows to test the equality of factor structures even in the presence of partly dissimilar stimuli across groups (Byrne, Shavelson, & Muthén, 1989).

Applications are straightforward from a statistical perspective. They are the only type of translations in which scalar equivalence of the total test score may be maintained. The use of statistical analyses such as t tests and analyses of variance to test the equality of means of cultural groups is meaningful only in the case of applications and if bias is completely absent. The possibility to carry out score comparisons is undoubtedly one of the main reasons for the popularity of applications. It should be acknowledged, however, that applications are highly restrictive in their assumptions : they require the absence of every type of bias. Even if a researcher chose the application option, it is recommended to routinely apply judgmental and psychometric methods of item bias detection to examine the appropriateness of the instrument. However, this practice does not yet safeguard against method and construct bias.

Sources of bias

The sources of bias in cross-cultural assessment are manifold and it is virtually impossible to present an exhaustive overview. The overview in Table I is based on a classification by Van de Vijver and Poortinga (1997) and shows the most typical sources for each of the three types of bias. A detailed list of examples for the different sources of bias can also be found in Van de Vijver and Leung (1997b).

Table I. *Typical Sources for the Three Types of Bias in Cross-Cultural Assessment (modified after Van de Vijver & Poortinga, 1997)*

Type of bias	Source of bias
Construct bias	- only partial overlap in the definitions of the construct across cultures - differential appropriateness of the behaviors associated with the construct (e.g., skills do not belong to the repertoire of one of the cultural groups) - poor sampling of all relevant behaviors (e.g., short instruments) - incomplete coverage of all relevant aspects/facets of the construct (e.g., not all relevant domains are sampled)
Method bias	- incomparability of samples (e.g., caused by differences in education, motivation)[a] - differences in environmental administration conditions, physical (e.g., recording devices) or social (e.g., class size)[b] - ambiguous instructions for respondents and/or guidelines for administrators[b] - differential expertise of administrators[b] - tester/interviewer/observer effects (e.g., halo effects)[b] - communication problems between respondent and tester/interviewer (including interpreter problems and taboo topics)[b] - differential familiarity with stimulus material[c] - differential familiarity with response procedures[c] - differential response styles (e.g., social desirability, extremity scoring, acquiescence)[c]
Item bias	- poor item translation and/or ambiguous items - nuisance factors (e.g., item may invoke additional traits or abilities) - cultural specifics (e.g., incidental differences in connotative meaning and/or appropriateness of the item content)

[a] Sample bias. [b] Administration bias. [c] Instrument bias.

Construct bias

Construct bias can occur if there is only partial overlap in the definitions of the construct across cultures. As mentioned previously, non-western societies´ conceptions of intelligence are often broader and usually include aspects such as social skills in addition to the primarily scholastic domains covered by intelligence tests developed according to Western concepts (e.g., Serpell, 1993 ; Super, 1983). In the personality area, Yang and Bond (1990) administered a set of emic (i.e., indigenous) Chinese descriptors together with a set of imported American descriptors to a group of Taiwanese subjects. Of the five Chinese factors identified, only four corresponded to the American factors. These results support the findings of Cheung et al. (1996) on indigenous Chinese personality dimensions such as “face” and “harmony.” Church (1987) also found indigenous personality constructs in the Filipino culture. In the field of value studies, the Chinese Culture Connection (1987) designed a value survey based entirely on Chinese values and administered it in 22 countries. It was found that only three out of the four factors were similar to those identified by Hofstede (1980), who used an instrument developed according to Western standards. The fourth factor, Confucian Work Dynamism, correlated highly with economic growth and was not covered by the Western concepts. In their review of selfhood in Eastern metapsychologies and philosophies, Hoshmand and Ho (1995) stressed the importance of social aspects in Chinese concepts as compared to the more individualistic concepts of an "autonomous self" in Western approaches ; a viewpoint shared by many other authors (e.g., Bochner, 1994 ; Paranjpe, 1995 ; Sampson, 1988).

Construct bias can also be caused by differential appropriateness of the behaviors associated with the construct in the different cultures. Kuo and Marsella (1977),

who studied Machiavellianism in China and the United States, argued that differences in "behavioral referents, correlates, and functional implications" (p. 165) question the equivalence of the construct in both countries. Another example is the study by Tanzer and Sim (1992 ; see also Tanzer, Sim, & Marsh, 1992). They used Corulla´s (1990) revised EPQ-Junior and found that Singaporean adolescents as compared to their British counterparts scored extremely high on the "Lie" scale (which is actually an indicator of social desirability). In Singapore, a society with high level of social engineering including heavy fines for littering and other forms of public misconduct, social conformity in this area is rather high. Thus, low endorsement rates on items like "Do you throw waste paper on the floor when there is no waste paper basket handy?" will reflect the degree of social conformity in the Singapore society (i.e., a cultural characteristic on the level of societies) rather than being an indicator of the response set phenomenon "social desirability" (i.e., a personality trait on the individual level).

Finally, poor sampling of all the relevant behaviors associated with the construct can also give rise to construct bias. Broad constructs are often represented by a relatively small number of items in a questionnaire or test and, thus, not all relevant domains are covered by the items. Embretson (1983) coined the term "construct underrepresentation" to refer to this insufficient sampling of all relevant domains. Short instruments can also result as consequence of the (necessary) removal of all biased item during test translations. In a cross-cultural Rasch analysis of the Cattell Culture Fair Intelligence Test between American and Nigerian students, Nenty and Dinero (1981) had to remove 24 out of 46 items because these items either did not fit the Rasch model or showed cross-cultural bias. With multidimensional self-report inventories in which the original scales usually consist of only a few items (e.g., 6-10), the problem of scale reduction is even more critical. In addition to poor sampling caused by instruments that are too short, incomplete coverage of all relevant aspects/facets associated with the construct can have similar effects.

Method bias

It is useful to distinguish three types of method bias, namely, sample bias, administration bias, and instrument bias. Sample bias or incomparability of samples occurs when the samples used differ in a variety of relevant characteristics ("nuisance factors") other than the target construct. Administration bias includes all sources of bias which are caused by the particular form of administration. Instrument bias subsumes all sources of method bias which are associated with the particular assessment instrument.

Comparisons of "remote" cultures (i.e., cultures which differ in many aspects) will often be characterized by sample incomparability because matching of samples in all relevant aspects is practically impossible to achieve. As a consequence of this sample bias, any observed cross-cultural differences can be attributed to the target construct as well as to the influence of "nuisance factors." In the case of cognitive tests, such nuisance factors could be differences in the educational system, novelty of the test situation, motivation of subjects, recruitment procedures, etc. Recruitment procedures, for example, are a rather underestimated source of sample bias in cognitive tests. In the U.S.A., studies are often conducted with students who are paid or given course credit points for their participation. Studies of other countries often employ undergraduates in psychology who participate out of curiosity or because they want to get some experience in how a test session is conducted. According to the "principle of effort justification" in cognitive dissonance theory, subjects who are "purely" volunteers should have higher levels of motivation and ego-involvement than subjects who receive sufficiently high reward. While higher motivation may result in more serious test-taking, there is ample evidence in the test anxiety literature that increased ego-involvement could - especially in the case of intelligence tests - cause ego-threatening thoughts which interfere with optimal task performance.

Administration bias can be caused by differences in the environmental administration conditions whether physical, technical, or social. In cross-cultural studies using technical equipment other then paper-and-pencil instruments, differential familiarity with the physical presence of measurement or recording devices (e.g., video cameras) could cause substantial cross-cultural differences in various non-target variables such as the subjects´ level of curiosity (caused by the novelty of the situation) or willingness to self-disclose. In a cross-cultural comparison of mental tests (cf. Tanzer, Gittler, & Ellis, 1995), subjects in one test location reported being disturbed by a freezing air-conditioned testing room (which was out of the experimenters´ control). Examples of social environmental conditions are individual versus group administration, amount of space between testees (in group testing), or class size (in educational settings). For example, primary school classes in Austria vary between 10 to 25 children and in Singapore between 30 to 45 children. In fields like sociometric peer status research (e.g., Asher & Coie, 1990 ; Newcomb, Bukowski, & Pattee, 1993), the validity of any cross-cultural comparison would suffer from such non-overlapping class sizes. Van de Vijver (1988, 1991) who tested inductive thinking in primary and secondary school children in Zambia, Turkey, and the Netherlands tried to solve the problem of differential class sizes by testing only half of the children per class in the countries with large class sizes. This approach, however, would not solve the problem in the above mentioned case of peer status research.

Administration bias can also be caused by ambiguous instructions for respondents and/or guidelines for administrators. In the case of differential expertise of administrators (e.g., senior faculty members versus undergraduate students), any ambiguity in the test instructions and/or guidelines can seriously threaten proper test administration. Thus, similar to the requirement of sample comparability discussed above, comparability of the administrators/interviewers in terms of general testing experience, familiarity with the specific test material used and pattern of tester-testee interaction is a further prerequisite for valid cross-cultural comparisons.

Tester/interviewer/observer effects such as obtrusiveness is another potential source of administration bias. The mere presence of a person from a different culture can strongly affect respondents' behavior (Singer & Presser, 1989). Super (1981) demonstrated that the presence of an observer may influence mother-child interactions. Likewise, several empirical studies have addressed the obtrusiveness of the test administrator´s culture in intelligence testing (cf. Jensen, 1980). Word (1977) found that White interviewers often placed African-American subjects on the defensive side by rephrasing or correcting their Black English. There exists also social-psychological and sociological research on interviewer effects which support a theory of deference. Subjects were more likely to display positive attitudes to a particular cultural group when they are interviewed by someone from that group (e.g., Cotter, Cohen, & Coulter, 1982 ; Reese, Danielson, Shoemaker, Chang, & Hsu, 1986). In general, however, the results on tester/interviewer/observer effects are quite inconsistent across studies (cf. Jensen, 1980 ; Singer & Presser, 1989).

A final source of administration bias are communication problems between the respondent and the tester/interviewer. Frequently, language problems are the reason for this source of bias because it is common in cross-cultural studies to carry out the test or interview in the second or third language of interviewers, respondents, or even both. Illustrations for such miscommunications between native and non-native speakers can be found in Gass and Varonis (1991). In addition, miscommunication in cross-cultural encounters may also arise from ethnocentric interpretations (e.g., Banks, Ge, & Baker, 1991; Barna, 1991; Cohen, 1987).

For instrument bias, differential familiarity with the stimulus material (e.g., the items) is a common source in cognitive tests. As mentioned previously, Deregowksi and Serpell (1971) found performance differences between Scottish and Zambian children in sorting photographs but not in sorting miniature models. When Western tests are administered to non-western cultural groups, differences in stimulus familiarity are almost certain. An often cited example is Piswanger´s (1975) cross-cultural comparison of the Viennese Matrices Test (Formann & Piswanger, 1979), a Raven-like figural inductive reasoning test. He compared the responses of (Arabic educated) Nigerian and Togolese high school students to those of the Austrian calibration sample. The most striking findings were cross-cultural differences in the item difficulties related to identifying and applying rules in horizontal direction (i.e., left to right). This was interpreted in terms of the different directions in writing Latin versus Arabic. Another example for differential stimulus familiarity are pictorial tests such as the Rosenzweig Picture-Frustration Test (Rosenzweig, 1977, 1978), the Preschool Symptom Self-Report (Martini, Strayhorn, & Puig-Antich, 1990), and the Pictorial Evaluation of Test Reactions (Toubiana, 1994). The items of these tests contain elements specific to a certain culture (e.g., Western style of dressing) and/or ethnic group (e.g., Causcasian faces).

Differential familiarity with the required response procedures can also be a source of instrument bias. A good illustration is the above-mentioned study of Serpell (1979) who asked Zambian and British children to reproduce a pattern using paper-and-pencil, plasticine, configurations of hand positions, and iron wire. Finally, differential response styles such as acquiescence and extremity ratings can cause method bias. A demonstration can be found in the work of Hui and Triandis (1989). These authors found that Hispanics tended to choose extremes on a five-point rating scale more often than did Anglo-Americans although no significant cross-cultural differences were found for ten-point scales. Ross and Mirowsky (1984) reported more acquiescence and socially desirable responses among Mexicans than among Anglo-Americans in a mental health survey. In a cross-cultural-comparison of the English version of Self-Description Questionnaire I (Marsh, 1988) between Australian and Singaporean adolescents, Tanzer, Sim, and Marsh (1994) found cross-cultural differences in the endorsement rates of competence items (as compared to interest items). These differences were attributed to the tendency to be modest which is still prevalent in Singapore. The two groups differed also substantially in their usage of the five rating scale categories (Tanner, 1995).

Not all sources of method bias discussed above are likely to affect all the five types of assessment procedures. The cross-classification in Table II is an attempt to indicate which sources of method bias can be expected to affect mental tests, questionnaires/inventories, observations, projective techniques, and interviews. While sample bias and most types of administration bias can be present in all five types of assessment procedures, communication problems are typically more prominent in interviews and, to a lesser degree, in projective techniques. In multicultural counseling (cf. McFadden, 1993 ; Paniagua, 1994 ; Ponterotto, Casas, Suzuki, & Alexander ; 1995 ; Wehrly, 1995) where counselor/therapist (i.e., interviewer) and client/patient (i.e., interviewee) are often from different cultural background, frictions in the communication process can easily be caused by insufficient familiarity with the client´s cultural background (e.g., taboo topics, Goodwin & Lee, 1994). Moreover, in the event that client and therapist do not speak the same language, the necessity to use an interpreter (who is almost never a trained psychologist, and more often than not, is just a bilingual without formal training as interpreter) will aggravate these problems even more.

Regarding instrument bias, almost by definition, the impact of the different sources will vary across the five assessment procedures. For mental tests, the possibility of bias caused by differential familiarity with the stimulus material and/or response procedures should be considered carefully. As for questionnaires and self-report inventories, the possibility of differential familiarity with the response procedure (e.g., usage of the rating scale categories in a Likert-type inventory) should also be given careful attention. As regards response styles, phenomena such as social desirability are more likely to affect studies using projective techniques, questionnaires, and interviews. On the other hand, subjects´ choice of his or her "preference point" on the speed-accuracy trade-off curve as a source of bias is only relevant in intelligence/aptitude tests.

Table II. *Prevalence of the Different Sources of Method Bias in the Five Types of Assessment Procedures*

Source of method bias	Assessment procedure				
	Mental tests	Questionnaires	Observations	Projective techniques	Interviews
Incomparability of samples[a]	+	+	+	+	+
Differences in environmental administration conditions (physical and/or social)[b]	+	+	+	+	+
Ambiguous instructions for respondents and/or guidelines for administrators[b]	+	+	+	+	+
Differential expertise of administrators[b]	+	+	+	+	+
Tester/interviewer/observer effects[b]	+	+	+	+	+
Communication problems between respondent and tester/interviewer[b]				+	+
Differential stimulus familiarity[c]	+			+	
Differential familiarity with response procedures[c]	+	+			
Differential response styles[c]	+	+		+	+

[a]Sample bias. [b]Administration bias. [c]Instrument bias.

Item bias

Although item bias can be produced by various sources, it is most frequently caused by poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, or influence of cultural specifics such as nuisance factors or connotations associated with the item wording. Poor item translation can either be caused by translation errors or by "genuine" linguistic idiosyncracies. Even translations that are linguistically correct may still have poor quality from a psychological point of view (e.g., the Swedish translation of "webbed feet" as "swimming feet" mentioned before). A common source of linguistic idiosyncracies is the metaphoric use of language. For example, English anxiety inventories often include the bodily symptom "getting butterflies in the stomach." In Dutch, however, the same metaphor ("Vlinders in je buik hebben") is often used to refer to a bodily symptom which occurs when someone falls in love and, thus, carries an erotic rather than a distress connotation. Another example of linguistic idiosyncracies is the well-known German term "Zeitgeist" which has no one-to-one English translation.

Ambiguities in the item content may trigger off different interpretations. Spielberger (1988) distinguished in his State-Trait Anger Expression Inventory (STAXI) three styles of anger expression, namely, Anger-out as expression of anger toward other people or objects, Anger-in as holding in or the suppression of angry feelings, and Anger-control as attempt to control the expression of anger. By and large, the postulated three-factor structure of the original English anger expression items was confirmed in several U.S. samples (e.g., Fuqua, Leonard, Masters, & Smith, 1991 ; Spielberger, 1988 ; Spielberger, Krasner, & Solomon, 1988), a sample of Singaporean Chinese (Tanzer, Sim, & Spielberger, 1996), as well as in translations into German (Schwenkmezger, Hodapp, & Spielberger, 1992), Italian (Spielberger & Comunian, 1992), Norwegian (Håseth, 1996), and Chinese (Yui Miao & Lin, 1990). However, in some of these studies, the item "I am secretly quite critical of others" shifted from Anger-in to Anger-out. Although anger is a universal emotion (cf. Mesquita & Frijda, 1992), expression of anger and coping with anger is governed by cultural factors. As such, the observed shifts could reflect genuine cross-cultural differences in the way Anger-out is expressed. For example, the need for "harmony" or "giving face" which is still prevalent in Chinese societies (e.g., Cheung et al., 1996) precludes any open confrontations which characterize Western ways of Anger-out expression. On the other hand, this item could be interpreted either as "holding grudges and not talking about it publicly" which would then be an Anger-in expression or as "talking negatively behind someone´s back" which would convey a covert Anger-out expression. In conclusion, the ambiguity in the interpretation of this item could be an alternative explanation for these shifts in factor loadings.

Item bias can also be caused by culture-specific nuisance factors such as the distance of the subject´s country to Poland in the item "What is the capital of Poland?". Finally, a

frequent source of item bias are cultural specifics in content and/or connotation of the item. The following example given by Szalay (1981) may serve as an illustration of culture-specific connotations.

The English word corruption corresponds beyond a shadow of a doubt to the Korean word pup´ae, but this does not ensure that the cultural meanings of the two words are the same. Different cultural experiences produce different interpretations not shown in conventional dictionaries. A systematic comparison of the Korean and American meanings of corruption shows that for both groups it involves negative, bad, improper behavior. An important difference is that in the American interpretation corruption is rejected on moral grounds ; it is wrong and it is a crime. For Koreans corruption is not morally wrong ; it is only bad in the sense that it interferes with the proper function of the government and social institutions ; and it is bad in its social consequences (p. 141).

An example of item bias caused by inappropriate item content is given by Van Haaften and Van de Vijver (1996). They had to remove the symptom "watched more television than usual" from a Western coping questionnaire when applied to Sahel dwellers without electricity in their homes.

Remedies

There is a rich literature on strategies to identify and deal with the three types of bias. Most strategies implicitly assume that bias is a nuisance factor that should be avoided. As a consequence, most of the literature are devoted to techniques that enable the reduction or even elimination of bias. It is less common to see bias as an indication of systematic cross-cultural differences that require further scrutiny. It is not until quite recently that the idea is gaining momentum that bias may yield important information about cross-cultural differences and can also be seen as a phenomenon that requires explanation (Poortinga & Van der Flier, 1989). A thorough discussion of all strategies proposed to deal with bias is far beyond the scope of the present chapter. We will restrict the presentation to the most salient techniques for addressing each of the three types of bias given in Table III.

Table III. *Strategies for Identifying and Dealing with Bias in Cross-Cultural Assessment*

Type of Bias	Strategies
Construct bias	- decentering (i.e., simultaneously developing the same instrument in several cultures) - convergence approach (i.e., independent within-culture development of instruments and subsequent cross-cultural administration of all instruments)
Construct bias and/or method bias	- use of informants with expertise in local culture and language - use samples of bilingual subjects - use of local surveys (e.g., content analyses of free-response questions) - non-standard instrument administration (e.g., "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 behaviour or test attitudes) - assessment of response styles - use of test-retest, training and/or intervention studies
Item bias	- judgmental methods of item bias detection (e.g., linguistic and psychological analysis) - psychometric methods of item bias detection (e.g., Differential Item Functioning analysis) - error or distracter analysis - documentation of "spare items" in the test manual which are be equally good measures of the construct as actually used test items

A powerful tool for examining construct bias is cultural decentering (Werner & Campbell, 1970). Words and concepts that are specific to one particular language or culture are eliminated. An example can be found in the study of Tanzer, Gittler, and Ellis (1995). Starting with a set of German intelligence/aptitude tests, they developed an English version of the test battery. Based on the results of pilots tests in Austria and the United States, both the German and English instructions and stimuli were modified before the main study was carried out. Another approach to deal with construct bias involves the convergence approach in which instruments are independently developed in different cultures and all instruments are then administered to subjects in all these cultures (Campbell, 1986).

Several techniques have been proposed to deal with both construct and method bias. One of these is the use of informants with a thorough knowledge of the local culture and language. Broer (1996) used a committee approach to develop the first version of a Spanish-language version of a test-attitude rating scales used in the Tanzer, Gittler, and Ellis (1995) study ; local informants were then asked to judge the accuracy of the instrument and to suggest necessary revisions.

In some studies an instrument is administered to a group of bilingual subjects. For example, Hocevar, El-Zahhar, and Gombos (1989) administered anxiety and arousability questionnaires to English-Hungarian bilinguals. Both language versions were used for all subjects. Even though many studies of equivalence involve bilinguals, their limitations must be acknowledged. First, bilinguals are usually not representative of the larger population because they are often better educated and have been more in contact with other cultures. Second, when carryover effects can be anticipated from the first to the second instrument such as in mental tests, it is important to control for order effects. In order to overcome these limitations, a combination of bilingual samples and monolingual samples would be useful. Sperber, Devellis, and Boehlecke (1994), in addition to combining the translation-backtranslation procedure with the committee approach, used both bilingual and monolingual samples to ensure the validity of their Hebrew version of an English-language survey on attitudes towards preventive medicine.

Another approach addressing both construct and method bias is the use of local surveys. For example, Super (1983) was interested in the question as to whether skills associated with intelligence were the same among the Kokwet, a farming community in Western Kenya as in Western countries. He found that in addition to reasoning, knowledge, and problem solving skills which are also found in Western studies, social aspects of intelligence such as knowing one's role in the family, ability to help members of the family, and obedience were also frequently mentioned. In another example, Mook, Oláh, Van der Ploeg, and Magnusson (1985) asked adolescents in the Netherlands, Hungary, India, Sweden, and Yemen to describe situations in which they have been afraid. The authors developed a classification scheme of anxiety-provoking situations in order to be able to compare the commonness of these situations in these countries. Brandt and Boucher (1986) were interested in the place of depression in emotion lexicons. They gathered emotion terms from informants in Australia, Indonesia, Japan, Korea, Malaysia, Puerto Rico, Sri Lanka, and the United States. A distinct depression cluster was found only in Japan, Indonesia, Sri Lanka, and the United States. For the other languages, depression-type words were predominantly subsumed by sadness clusters. A well-known example in the field of self-concept research is Kuhn and McPartland´s (1954) "Twenty Statements Test." Instead of asking the subjects to rate a series of given self-describing statements, the open response format of this sentence completion test ("I am ...") allows a picture of the most salient aspects of the subjects´ self-definitions to be drawn.

When an instrument is administered for the first time in a cultural group, a pilot study in which the instrument is administered in a nonstandard way may provide useful information. In such an administration the subject is encouraged to indicate how he or she interprets the stimuli (instruction and stimuli) and to motivate responses. In this way, a researcher gets a good insight into the face validity of the instrument and the administration. Poorly formulated questions or other problems with the instrument may be quickly revealed. After the pilot study, in the actual data collection, deviations of the standard administration can also be useful. For example, the Viennese Matrices Test (WMT ; Formann & Piswanger ; 1979) was administered to freshmen in Chile and Austria (Broer, 1996 ; Tanzer, Gittler, & Ellis, 1995). The manual specifies a total testing time of 25 minutes which is sufficient for most subjects in Austria (where the test was developed) to complete the task. This time limit was lifted in the cross-cultural study in order to ensure that all subjects would have ample testing time. It was found that over 90% of the Austrian subjects completed the test in 25 minutes while in Chile only 55% did so. The average test scores obtained with an unlimited testing time did not differ significantly. However, the cross-cultural differences obtained under standard instructions might have been significant, thereby incorrectly indicating that the groups differ in inductive reasoning skills.

"Thinking aloud," a technique regularly employed in cognitive psychology because it gives insight into the cognitive processes involved in the solution processes of mental tasks, may turn out to be useful in cross-cultural comparison of mental tests too. Using this technique, Putz-Osterloh (e.g., Putz-Osterloh & Lüer, 1979) demonstrated that in the "cube-comparison" subscale of a widely-used German intelligence test battery (IST-70 ; Amthauer, 1970) a certain type of cube-comparison items can by solved by a much easier non-spatial strategy. It needs, however, a certain level of technical education to be aware of this strategy. Based on this result, Gittler (1990) excluded this type of items from the Three-dimensional Cubes Test (3DC), a Rasch-calibrated spatial ability test. The new test proved its favorable psychometric properties both in a large-scale Austrian calibration study (Gittler, 1990) as well as in several cross-cultural studies (Tanzer, Gittler, Ellis, 1995 ; Tanzer, Gittler, & Sim, 1994).

Another technique that addresses both construct and

method bias is the cross-cultural comparison of nomological networks which can be carried out by different methods. For instance, monotrait-multimethod studies can be conducted to examine the (lack of) consistency of cross-cultural findings across measurement methods (cf. Serpell's, 1979, study of perceptual skills). Another way is to examine the convergent and discriminant validity of the target construct. This method was frequently used during the development of the WISC(R) and the WAIS(R) adaptations. Tanzer, Sim, and Marsh (1992), however, cautioned against judging the presence of construct (or method) bias solely on the basis of cross-cultural differences found in the nomological network. Nomological network analysis involves other constructs besides just the target construct and, thus, any cross-cultural differences found can either be caused by the target construct or by the other constructs. For example, Tanzer and Sim (1991, 1997) found that in Singapore good students worry more about their performance during tests than weak students whereas the contrary was found in most other test anxiety research. For the other components of test anxiety (i.e., tension, low confidence, and cognitive interference), no cross-cultural differences were found. The authors attributed the inverted worry-achievement relationship to characteristics of the educational system, especially the "kiasu" (fear of losing out) syndrome which is deeply entrenched in the Singaporean society, rather then to construct bias in the internal structure of test anxiety.

Sometimes, a construct can only be expressed by a particular word or phrase. If such a key phrase is of pivotal importance as was the case with the word "injustice" in a study by Mikula, Petri, and Tanzer (1990) on everyday experiences of injustice, culture-specific connotations of that key phrase would seriously threaten the validity of any cross-cultural comparison. Thus, the connotative equivalence of key phrases and their translations should be carefully investigated with methods such as free word associations or semantic differentials (e.g., Szalay, 1981) .

There are also strategies that mainly address method bias. A first proposal involves the extensive training of administrators. Such training is not unique to cross-cultural research (e.g., Gutjahr, 1985) even though its need may be more pressing. Extensive training is required in order to ensure that interviewers will administer an interview in the same way in various cultural groups. If the cultures of the interviewer and the interviewee differ (which is common in studies involving multicultural groups), it is imperative to make the interviewers aware of the relevant cultural specifics such as taboo topics (e.g., Goodwin & Lee, 1994). As mentioned earlier, the issue of cross-cultural training is well-covered in the literature on multi-cultural counseling (cf. McFadden, 1993 ; Paniagua, 1994 ; Ponterotto, Casas, Suzuki, & Alexander ; 1995 ; Wehrly, 1995). Further aspects which are relevant for a cross-cultural training of interviewers can be found in the literature on intercultural communication and communication competence (e.g., Asante & Gudykunst, 1989 ; Cohen, 1987 ; Coupland, Giles, & Wiemann, 1991 ; Fiedler, Mitchell, & Triandis, 1971 ; Landis & Bhagat, 1996 ; Multicultural Training, 1994 ; Schneller, 1989 ; Ting, Toomey & Korzenny, 1991).

A related approach amounts to the development of a detailed manual and administration protocol. The manual should ideally specify the test or interview administration and should describe contingency plans on how to intervene in common administration problems, as was done in the above mentioned spatial ability test (Gittler, 1990). In studies which use open response formats, a detailed scoring system should be provided. Particularly, detailed instructions in mental tests should clearly describe what is expected of the subject. Moreover, sufficient examples and exercises should be provided in order to guarantee an unambiguous communication of what the subject should do and can expect. The role of exercises is illustrated in a study by Tanzer, Gittler, and Sim (1994). They applied a spatial ability task to Austrian and Singaporean youngsters. The authors found a good fit of the data to the Rasch model after the elimination of the first item which was unexpectedly difficult in the Singaporean group. The authors attributed the observed cross-cultural difference in the difficulty of the first item to differential warming-up effects. Helms-Lorenz and Van de Vijver (1995) administered a computerized cognitive test to native and migrant children in the Netherlands. In order to make sure that the instruction is correctly understood, the subjects have to solve various examples. If they make errors, the exercises are presented again. The actual test session starts when they have correctly solved all exercises.

An entirely different approach of method bias is the use of subject and context variables. In cross-cultural studies, it is usually impossible to match cultural groups on all variables that are relevant to the variable under study. For example, in mental testing, differences in educational background may be a confounding variable. Poortinga and Van de Vijver (1987) proposed to include these confounding variables into the design of a study. When a confounding variable has been measured (e.g., a measure of the educational background of all subjects), it becomes possible to statistically check its influence in a covariance or hierarchical regression analysis. Poortinga (cf. Poortinga & Van de Vijver, 1987) studied the habituation of the orienting reflex among illiterate Indian tribes and Dutch conscripts. The skin conductance response, the dependent variable, was significantly larger in the Indian group. It could be argued that intergroup differences in arousal could account for these differences. Arousal was operationalized as the spontaneous fluctuations in skin conductance response in a control condition. After statistically controlling for these fluctuations using a hierarchical regression analysis, the cross-cultural differences in habituation of the orienting reflex disappeared.

Some instruments will allow for the use of collateral information that provides evidence about the presence or absence of method bias. Thus, the administration time of power tests can be used to examine cross-cultural differences in response time, as was illustrated in the study mentioned above (Broer, 1996 ; Tanzer, Gittler, & Ellis, 1996). Similarly, Verster (1983) examined the performance of adult Whites and Blacks in South Africa on various mental tests, ranging from tapping tasks to inductive reasoning. The tests were administered without time limit by a computer. The use

of the computer enabled him to study both the speed and the accuracy of the responses. Using structural equation modeling, separate models were fitted to the speed and accuracy data. The speed data required a more complex model than the accuracy data. The cross-cultural differences in the fitted model were larger for the accuracy than for the speed data. This study demonstrates the influence of the method of data collection on the outcome of a study ; the collateral information, the speed measures, did not show the same pattern of cross-cultural differences than the accuracy measure which is commonly used in power tests.

There are also examples of studies that measure various test-taking dispositions such as motivational factors (e.g., Arvey, Strickland, Drauden, & Martin, 1990 ; Schmit & Ryan, 1992) or performance-debilitating levels of test anxiety (e.g., Ball, 1995). Oakland, Gulek, and Glutting (1996) assessed test-taking behaviors among Turkish children and their results, similar to those obtained with American children, showed that these behaviors are significantly correlated with the WISC-R IQ. Arvey, Strickland, Drauden, and Martin (1990), working with adults, found significant Black-White differences on test taking attitudes ; Whites reported to be more motivated to exert effort and work hard while Blacks scored higher on preparation.

In addition to test-taking dispositions, one can also manipulate or measure response styles. As described before, Hui and Triandis (1989) found that the number of alternatives in a Likert scale can influence the measurement outcome. Cross-cultural research with the Eysenck Personality Questionnaire has provided ample evidence for fairly systematic cross-cultural differences in social desirability (e.g., Eysenck & Abdel-Khalek, 1989 ; Eysenck & Kozeny, 1990 ; Eysenck, Makaremi, & Barrett, 1994 ; Eysenck, & Renner, 1987 ; Eysenck & Tambs, 1990 ; Eysenck &Yanai, 1985 ; Sanderman, Eysenck, & Arrindell, 1991 ; Tanzer, Sim, & Marsh, 1992). Moreover, there are indications that social desirability varies systematically with sample characteristics. Ross and Mirowsky (1984) administered an adapted version of the Marlowe-Crowne Social Desirability Scale to Anglo-American, Mexican, and Mexican-American adults, aged 18-65 years, in face-to-face interviews. As mentioned earlier, they found that people of Mexican origin and people of lower socioeconomic status tended to show more acquiescence and social desirability. The authors interpreted their data as support for a model that "the practices of giving socially desirable responses and agreeing with statements regardless of their content are more common in social groups that are relatively powerless" (p. 189). However, empirical scrutiny is required to examine whether the practices are due to powerlessness, lower education, or some other related source.

Evidence on the presence of method bias can also be collected by applying test-retest, training, or intervention studies. Patterns of pretest-posttest change that are different across cultures point to the presence of method bias. Van de Vijver, Daal, and Van Zonneveld (1986) administered a short training of inductive reasoning to primary school pupils from the Netherlands, Surinam, and Zambia. The gain patterns were not identical across groups. The Zambian subjects showed considerable score increments, both in the experimental and in an untrained control group. The differential gain pattern was interpreted as evidence of method bias. Similarly, Foorman, Yoshida, Swank, and Garson (1989) administered a training in solving computerized geometric matrices to American and Japanese pupils. No differences were found at the pretest whereas at the posttest the error decrease was more accompanied by reaction time increase for the American than for the Japanese children. The finding was attributed to the "Japanese children's expeditious style of considering task-related information" (p. 295).

There are two kinds of procedures to assess item bias : judgmental procedures, either linguistic or psychological, and psychometric procedures. An example of a linguistic procedure can be found in Grill and Bartel (1977). They examined the Grammatic Closure subtest of the Illinois Test of Psycholinguistic Abilities for bias against speakers of nonstandard English. In the first stage, potentially biased items were identified. Error responses of American Black and White children indicated that more than half of the errors on these items were accounted for by responses that are appropriate in nonstandard English. Examples of psychometric procedures are numerous. Valencia, Rankin, and Livingston (1995) examined the item bias of the Mental Processing Scales and the Achievement Scale of the Kaufman Assessment Battery for Children with Mexican American and White pupils. Using a partial correlation index (that controlled for age, sex, and ability), the authors found 17 of 120 items of the first scale and 58 of 92 items of the last scale to be biased. With respect to the latter test, it is questionable whether the remaining 34 items will constitute an appropriate instrument that still measures the same construct as the full scale. It is quite likely that, in the terminology of the present article, the scale is plagued by construct and/or method bias. 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. A replication study with a new U.S. sample showed that 6 of the 11 biased items were again biased.

Finally, the method of error or distracter analysis could be a promising approach for cross-cultural comparisons of mental tests with multiple-choice items. This approach identifies typical types of errors and, with carefully planned distracters, give insight into the cognitive processes involved in the solution process. For example, Vodegel Matzen, Van der Molen, and Dudink (1994) used such an analysis to demonstrate that errors in the Standard Progressive Matrices were often caused by omitting a rule. Unfortunately, like the method of "thinking aloud," error analyses have hardly been applied in cross-cultural research.

Although linguistic and psychometric procedures of item bias analyses can often identify items which are merely biased because of cultural specifics such as idiomatic use of language, they usually cannot solve the problem. In order to avoid the problem of construct underrepresentation caused by too short tests, new items must be substituted for biased

items. Manuals of the original tests including ample documentation of "spare items" which are as good measures of the construct as actually used test items can provide helpful information for substituting out biased items.

On the other hand, test adaptation manuals or reports should provide detailed documentation of all changes done, along with (linguistic and/or psychometric) justification. This principal is also included in the recently developed "Guidelines for Test Translations" of the International Test Commission (cf. Van de Vijver & Hambleton, 1996, Guideline #19). An example of a well-documented test adaptation is Håseth´s (1996) report on the Norwegian translation of the STAXI (Spielberger, 1988).

Conclusion

It cannot be taken for granted that scores obtained in one culture can be compared across cultural groups. Score differences observed in cross-cultural comparisons may have a partly or entirely different meaning than those in intracultural comparisons. If this is the case, bias is said to occur. Therefore, bias and its counterpart, equivalence, are essential concepts in cross-cultural research. A distinction in three types of bias had been made, depending on whether the source of bias is located at the level of the construct (labeled construct bias), instrument administration (method bias), or the separate items (item bias). The origin, identification, and ways of dealing with bias were discussed.

We concur with the view that intergroup comparisons in studies that do not address bias are often unable to unambiguously interpret observed differences and to rule out alternative explanations such as intergroup differences in stimulus familiarity or response styles. Yet, in the design of empirical studies it is often possible to be very selective in considering the choice of alternative explanations. A careful review of the literature will often reveal the types of bias to be expected in a particular field of cross-cultural research and/or a particular assessment technique. Regarding method bias, the likelihood that all the different sources will threaten the cross-cultural validity of mental tests, questionnaires, inventories, observations, projective techniques, and interviews is extremely low. A careful choice of assessment instrument can help to control method bias. Moreover, the likelihood of a certain type of bias will not just depend on the type of research question but also on the cultural distance of the groups involved, defined here as a generic term for all aspects in which the groups differ and which are relevant to the target variable. All these aspects can, in principle, statistically explain observed cross-cultural score differences. Obviously, the more aspects, the more bias threats.

Another important consideration is the research question of a study and the type of equivalence aimed at. A distinction can be made between structure- and level-oriented studies (Van de Vijver & Leung, 1997a, 1997b). In structure-oriented studies, the identity of psychological constructs is addressed. For example, the question as to whether the Big Five constitute an adequate description of personality in various cultural groups can be asked (cf. *European Review of Applied Psychology*, 1994, 44, 1). Similarly, the universality of structures underlying mental test performance can be examined. In structure-oriented studies, one will usually not be concerned with method bias. As score averages will not be compared across cultures, all sources of bias that exert a more or less uniform influence across all items will not challenge the validity of the outcome. Likewise, as the establishment of the identity of mean structures is not the aim of structure-oriented studies, item bias does not challenge the validity of intergroup comparisons in these studies either.

On the other hand, level-oriented studies examine differences in averages across cultural groups. For instance, are Chinese more introvert than British? Bias requires more scrutiny in level-oriented studies. If a study uses a design in which intracultural differences are compared across cultures such as in pretest-posttest designs, or in comparisons of gender differences across cultures, measurement unit level equivalence suffices ; as a consequence, method bias will usually not jeopardize the findings. It is only when bias sources affect intracultural comparisons differentially that method bias threatens the validity of the conclusions. An example would be a training study in which the children in one culture learn different aspects from training such as test-wiseness. On the other hand, if the aim of a study is full scale equivalence (e.g., a cross-cultural comparison of average scores), all forms of bias will form a threat to the validity of the inferences.

The intricacies in the establishment of numerical score comparisons may have an interesting analogy in the methodological problems in the measurement of change. Like a statistically adequate measure of change, full scale equivalence is difficult to establish. In view of all problems in change measurement, Cronbach and Furby (1970) argued that in many cases we may not be at all interested in establishing change and that we can often rephrase our research question in such a way that there is no need to assess change. Analogously, it could be argued that the importance of the comparison of average scores across cultures is overrated. An observation of cross-cultural mean score differences is interesting, but full-fledged cross-cultural research will go beyond the exploration of averages of various cultural groups and will try to look for a valid explanation. The observation of cross-cultural differences should only be the vehicle for further scrutiny. Our interest in cross-cultural research is often not in the establishment of cross-cultural equivalence but in the fact that observed cross-cultural differences provide a starting point for further research, addressing questions such as causes and patterning of these differences. In studies of patterning, the objective is to find systematic patterns in cross-cultural differences. For example, Bond and Smith (1996) analyzed cross-cultural and historical changes in conformity as measured in the Asch-type line judgment task. In a meta-analysis, using 133 studies from 17 countries, the authors were able to show that in the U.S. conformity has declined since the 1950s and the individualism score of a country was significantly correlated with conformity.

An interesting way to avoid the intricacies of establishing scalar equivalence is to search for cross-cultural similarities in psychological functioning. Paradoxical as this may sound

(after all, cross-cultural research owes its existence to differences), a study of cross-cultural similarities, if properly planned, can yield valuable insights. The underlying rationale is simple : Similarities across cultures usually do not require an explanation and can be interpreted at face value whereas differences are usually open to multiple interpretations and require additional explanations. At first sight, one may want to argue that results indicating that two highly similar cultures do not differ significantly with regard to some psychological construct is not very exciting. However, this argument does not hold at all when the cultures differ in many aspects. The more differences on extraneous variables there are between cultures, the more valuable the finding that cultures do not differ on some target variable. All aspects in which the cultures differ can then be assumed to have no influence on the target variable.

When new tests are developed, it becomes customary to report data on their reliability and validity. We welcome and encourage a similar development in cross-cultural research vis-à-vis bias and equivalence. In the beginning era of cross-cultural research, the implicit agenda guiding much research was the demonstration of cross-cultural differences. Cross-cultural research has developed beyond the stage of a mere demonstration of cross-cultural differences into a new era in which the interpretation of these differences is pivotal. Because bias and equivalence are central concepts in the interpretation, their impact can be expected to grow.

REFERENCES

AMTHAUER, R. (1970) *Intelligenze-Struktur-Test IST* [The Intelligence-Structure-Test IST]. Göttingen, Germany : Hogrefe.

ARVEY, R. D., STRICKLAND, W., DRAUDEN, G., & MARTIN, C. (1990) Motivational components of test taking. *Personnel Psychology*, 43, 695-716.

ASANTE, M. K., & GUDYKUNST, W. B. (Eds.) (1989) *Handbook of international and intercultural communication*. London : Sage.

ASHER, S. R., & COIE, J. D. (Eds.) (1990) *Peer rejection in childhood*. Cambridge : Cambridge University Press.

BALL, S. (1995). Anxiety and test performance. In C. D. Spielberger & P. R. Vagg (Eds.), Test anxiety. *Theory, assessment, and treatment* (pp. 107-113). Washington, DC : Taylor & Francis.

BANKS, S. P., GE, G., & BAKER, J. (1991) Intercultural encounters and miscommunication. In N. Coupland, H. Giles, & J. M. Wiemann (Eds.), *Miscommunication and problematic talk* (pp. 103-120). Newbury Park, CA : Sage.

BARNA, L. M. (1991) Stumbling blocks in intercultural communication. In L. A. Samovar & R. E. Porter (Eds.), *Intercultural communication : A reader (6th ed.)*. Belmont, CA : Wadsworth.

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

BOCHNER, S. (1994) Cross-cultural differences in the self-concept. A test of Hostede´s individualism/collectivism distinction. *Journal of Cross-Cultural Psychology*, 25, 273-283.

BOND, R., & SMITH, P. B. (1996) Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, 119, 111-137.

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.

BRANDT, M. E., & BOUCHER, J. D. (1986) Concepts of depression in emotion lexicons of eight cultures. *International Journal of Intercultural Relations*, 10, 321-346.

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

BRISLIN, R. W. (1986) The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Newbury Park, CA : Sage.

BROER (1996) *Rasch-homogene Leistungstests (3DW, WMT) im Kulturvergleich Chile-Österreich. Erstellung einer spanischen Version einer Testbatterie und deren interkulturelle Validierung in Chile* [Cross-cultural comparison of the Rasch-calibrated tests 3DW and WMT between Chile-Austria and the development of a Spanish version of the test battery]. Unpublished master´s thesis, University of Vienna, Austria.

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.

BYRNE, B. M., SHAVELSON, R. J., & MUTHÉN, B. (1989) Testing for the equivalence of factor covariance and mean structures : The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466.

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

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.

CHINESE CULTURE CONNECTION (1987) Chinese values and the search for culture-free dimensions of culture. *Journal of Cross-Cultural Psychology*, 18, 143-164.

CHURCH, T. A. (1987) Personality research in a non-Western setting : The Philippines. *Psychological Bulletin*, 102, 272- 292.

CORULLA, W. J. (1990) A revised version of the Psychoticism Scale for children. *Personality and Individual Differences*, 11, 65-76.

COTTER, P. R., COHEN, J., & COULTER, P. (1982) Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, 46, 278-284.

COHEN, R. (1987) International communication : An intercultural approach. *Cooperation and Conflict*, 22, 63-80.

COUPLAND, N., GILES, H., & WIEMANN, J. M. (Eds.). (1991). *Miscommunication and problematic talk*. Newbury Park, CA : Sage.

CRONBACH, L. J., & FURBY, L. (1970) How should we measure "change" -- or should we? *Psychological Bulletin*, 74, 68-80.

CRONBACH, L. J., & MEEHL, P. E. (1955) Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

DEREGOWSKI, J. B., & SERPELL, R. (1971) Performance on a sorting task : A ross-cultural experiment. *International Journal of Psychology*, 6, 273-281.

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

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.

EUROPEAN REVIEW OF APPLIED PSYCHOLOGY (1994) The Big Five. *ERAP Special Issue*, 44, n°1.

EYSENCK, S. B. G., & ABDEL-KHALEK, A. M. (1989) A cross-cultural study of personality : Egyptian and English children. *International Journal of Psychology*, 24, 1-11.

EYSENCK, S. B. G., & KOZENY, J. (1990) Cross-cultural comparisons of personality : Czech and English subjects. *Studia Psychologica*, 32, 255-259.

EYSENCK, S. B. G., MAKAREMI, A., & BARRETT, P. T. (1994) A cross-cultural study of personality : Iranian and English children. *Personality and Individual Differences*, 16, 203-210.

EYSENCK, S. B. G., & RENNER, W. (1987) A cross-cultural comparison of personality : English and Austrian children. *European Journal of Personality*, 1, 215-221.

EYSENCK, S. B. G., & TAMBS, K. (1990) Cross-cultural comparison of personality: *Norway and England. Scandinavian Journal of Psychology*, 31, 191-197.

EYSENCK, S. B. G., & YANAI, O. (1985) A cross-cultural study of personality : *Israel and England. Psychological Reports*, 57, 111-116.

FIEDLER, F. E., MITCHELL, T., & TRIANDIS, H. C. (1971) The cultural assimilator : An approach to cross-cultural training. *Journal of Applied Psychology*, 55, 95-102.

FOORMAN, B. R., YOSHIDA, H., SWANK, P. R., & GARSON, J. (1989) Japanese and American children's styles of processing figural matrices. *Journal of Cross-Cultural Psychology,* 20, 263-295.

FORMANN, A. K., & PISWANGER, K. (1979) *Wiener Matrizen-Test. Ein Rasch-skalierter sprachfreier Intelligenztest* [The Viennese Matrices Test. A Rasch-calibrated non-verbal intelligence test]. Weinheim, Germany : Beltz Test.

FUQUA, D. R., LEONARD, E., MASTERS, M. A., & SMITH, R. J. (1991) A structural analysis of the State-Trait Anger Expression Inventory. *Educational and Psychological Measurement*, 51, 439-446.

GASS, S. M., &VARONIS, E. M. (1991) Miscommunication in nonnative speaker discourse. In N. Coupland, H. Giles, & J. M. Wiemann (Eds.), *Miscommunication and problematic talk* (pp. 121-145). Newbury Park, CA: Sage.

GEISINGER, K. F. (1994) Cross-cultural normative assessment : Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6, 304-312.

GITTLER, G. (1990) *3DW. Dreidimensionaler Würfeltest. Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens.Theoretische Grundlagen und Manual* [The Three-dimensional Cube Test 3DC. A Rasch-calibrated spatial ability test. Theoretical background and test manual]. Weinheim, Germany : Beltz Test.

GOODWIN, R. & LEE, I. (1994) Taboo topics among Chinese and English friends. A cross-cultural comparison. *Journal of Cross-Cultural Psychology*, 25, 325-338.

GRILL, J. J., & BARTEL, N. R. (1977) Language bias in tests : ITPA grammatic closure. *Journal of Learning Disabilities*, 10, 229-235.

GUTJAHR, G. (1985) *Psychologie des Interviews in Praxis und Theorie* [Psychology of interviews. Theory and praxis] Heidelberg, Germany : Sauer.

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, the Netherlands : Kluwer.

HAMBLETON, R. K., SWAMINATHAN, H., & ROGERS, H. J. (1991) *Fundamentals of item response theory*. Newbury Park, CA : Sage.

HÅSETH, K. J. (1996) The Norwegian adaptation of the State-Trait Anger Expression Inventory. In C. D. Spielberger & I. Sarason (Eds.), *Stress and Emotion* (Vol. 16, pp. 83-106). Washington : Francis & Taylor.

HELMS-LORENZ, M., & VAN DE VIJVER, F. J. R. (1995) Cognitive assessment in education in a multicultural society. *European Journal of Psychological Assessment*, 11, 158-169.

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

HOCEVAR, D., EL-ZAHHAR, N., & GOMBOS, A. (1989) Cross-cultural equivalence of anxiety measurements in English-Hungarian bilinguals. In R. Schwarzer, H. M. Van der Ploeg, & C. D. Spielberger (Eds.), *Advances in test anxiety research* (Vol. 6, pp. 223-231). Lisse, the Netherlands : Swets.

HOF STEDE, G. (1980) *Culture's consequences: International differences in work-related values*. Beverly Hills, CA : Sage Publications.

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

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

HOSHMAND, L. T., & HO, D. Y. F. (1995) Moral dimensions of selfhood : Chinese traditions and cultural change. *World Psychology*, 1, 47-69.

HUI, C. H., & TRIANDIS, H. C. (1989) Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20, 296-309.

JENSEN, A. R. (1980) *Bias in mental testing*. New York : Free Press.

KAGITCIBASI, C. (1996) *Family and human development across cultures. A view from the other side*. Hillsdale, NJ : Erlbaum.

KLIEME, E., & STUMPF, H. (1991) DIF : A computer program for the analysis of differential item performance. *Educational and Psychological Measurement*, 51, 669-671.

KUHN, M. H. & MCPARTLAND, T. (1954) An empirical investigation of self-attitudes. *American Sociological Review*, 19, 68-76.

KUO, H. K., & MARSELLA, A. J. (1977) The meaning and measurement of Machiavellianism in Chinese and American college students. *Journal of Social Psychology*, 101, 165-173.

LANDIS, D., & BHAGAT, R. S. (Eds.). (1996) *Handbook of intercultural training (2nd ed.)*. London : Sage.

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.

MARSH, H. W. (1988) *Self-Description-Questionnaire I . SDQ-I manual and research monograph*. San Antonio, TX : The Psychological Corporation.

MARTINI, D. R., STRAYHORN, J. M., & PUIG-ANTICH, J. (1990) A symptom self-report measure for preschool children. *Journal of the American Academy of Child and Adolescent Psychiatry,* 29, 594-600.

MCFADDEN, J. (Eds.). (1993) *Transcultural counseling : Bilateral and international perspectives*. Alexandria, VA : American Counseling Association.

MERCER, J. R. (1984) What is a racially and culturally nondiscriminatory test? In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 293-356). New York : Plenum.

MESQUITA, B., & FRIJDA, N. H. (1992) Cultural variations in emotions : A review. *Psychological Bulletin*, 112, 179-204.

MIKULA, G., PETRI, B., & TANZER, N. (1990) What people regard as unjust : Types and structures of everyday experiences of injustice. *European Journal of Social Psychology,* 20, 133-149.

MOOK, J., OLÁH, A., VAN DER PLOEG, H. M., & MAGNUSSON, D. (1985) Culture, age and sex as moderating factors for expected consequences in achievement-demanding and socially evaluative situations. In H. M. Van der Ploeg, R. Schwarzer, & C. D. Spielberger (Eds.), *Advances in test anxiety research* (Vol. 4, pp. 169-182). Lisse, the Netherlands : Swets.

MULTICULTURAL TRAINING. (1994) *The Counseling Psychologist*, 22(2) [Special Issue].

NENTY, H. J., & DINERO, T. E. (1981) A cross-cultural analysis of the fairness of the Cattell Culture Fair Intelligence Test using the Rasch model. *Applied Psychological Measurement*, 5,355-368.

NEWCOMB, A. F., BUKOWSKI, W. M., & PATTEE, L. (1993) Children´s peer relations : A meta-analytic review of popular, rejected, neglected, controversial, and average sociometric status. *Psychological Bulletin*, 113, 99-128.

OAKLAND, T., GULEK, C., & GLUTTING, J. (1996) Children's test-taking behaviors : A review of literature, case study, and research of children. *European Journal of Psychological Assessment* (Bulletin of the International Test Commission), 12, 240-246.

PANIAGUA, F. A. (1994) *Assessing and treating culturally diverse clients : A practical guide*. Thousand Oaks, CA : Sage.

PARANJPE, A. C. (1995) The denial and affirmation of self : The complementary legacies of East and West. *World Psychology*, 1, 9-46.

PISWANGER, K. (1975) *Interkulturelle Vergleiche mit dem Matrizentest von Formann* [Cross-cultural comparisons with Formann's Matrices Test]. Unpublished doctoral dissertation, University of Vienna, Vienna.

PONTEROTTO, J. G., CASAS, J. M., SUZUKI, L. A., & ALEXANDER, C. M. (Eds.). (1995) *Handbook of multicultural counseling*. Thousand Oaks, CA : Sage.

POORTINGA, Y. H. (1989) Equivalence of cross cultural data : An overview of basic issues. *International Journal of Psychology*, 24, 737-756.

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

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.

POORTINGA, Y. H., & VAN DER FLIER, H. (1988) The meaning of item bias in ability tests. In S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context* (pp. 166-183). Cambridge : Cambridge University Press.

PUTZ-OSTERLOH, W., & LÜER, G. (1979) Wann produzieren Probanden räumliche Vorstellungen beim Lösen von Raumvorstellungsaufgaben [When do subjects use spatial strategies for solving spatial ability items]? *Zeitschrift für experimentelle und angewandte Psychologie*, 26, 138-156.

REESE, S. D., DANIELSON, W. A., SHOEMAKER, P. J., CHANG, T., & HSU, H.-L. (1986) Ethnicity-of-interviewer effects among Mexican-Americans and Anglos. *Public Opinion Quarterly*, 50, 563-572.

ROSENZWEIG, S. (1977) *Manual for the children´s form of the Rosenzweig Picture-Frustration (P-F) Study*. St. Louis : Rana House.

ROSENZWEIG, S. (1978) *The Rosenzweig Picture-Frustration (P-F) Study : Basic*

manual. St. Louis : Rana House.

ROSS, C. E., & MIROWSKY, J. (1984) Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior,* 25, 189-197.

SAMPSON, E. (1988) The debate on individualism : Indigenous psychologies of the individual and their role in personal and societal functioning. *American Psychologist*, 43, 15-22.

SANDERMAN, R., EYSENCK, S. B. G., & ARRINDELL, W. A. (1991) Cross-cultural comparisons of personality, The Netherlands and England. *Psychological Reports*, 69, 1091-1096.

SCHMIT, M. J., & RYAN, A. M. (1992) Test-taking dispositions : A missing link ? *Journal of Applied Psychology*, 77, 629-637.

SCHNELLER, R. (1989) Intercultural and intrapersonal processes and factors of misunderstanding : Implications for multicultural training. *International Journal of Intercultural Relations*, 13, 465-484.

SCHWENKMEZGER, P., HODAPP, V., & SPIELBERGER, C. D. (1992*) Das State-Trait-Argerausdrucks-Inventar STAXI* [The German adaptation of the State-Trait Anger Expression Inventory]. Bern : Huber.

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

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

SHEPARD, L., CAMILLI, G., & AVERILL, M. (1981) Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317-375.

SINGER, E., & PRESSER, S. (1989) The interviewer. In E. Singer & S. Presser (Eds.), *Survey research methods* (pp. 245-246). Chicago : University of Chicago Press.

SPERBER, A. D., DEVELLIS, R. F., & BOEHLECKE, B. (1994) Cross-cultural translation. Methodology and validation. *Journal of Cross-Cultural Psychology*, 25, 501-524.

SPIELBERGER, C. D. (1988) *State-Trait Anger Expression Inventory research edition. Professional manual*. Odessa, FL : Psychological Assessment Resources.

SPIELBERGER, C. D., & COMUNIAN, A. L. (1992*) STAXI. State-Trait Anger Expression Inventory. Versione e adattamento italiano a curi di Anna Laura Comunian. Manuale* [Test manual of the Italian version of the State-Trait Anger Expression Inventory] : Firenze : Organizzazioni Speziali.

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.

SPIELBERGER, C. D., KRASNER, S. S., & SOLOMON, E. P. (1988) The experience, expression, and control of anger. In M. P. Janisse (Ed.), *Health psychology : Individual differences and stress (*pp. 89-108). New York : Springer.

SUPER, C. M. (1983) Cultural variation in the meaning and uses of children's "intelligence." In J. B. Deregowski, S. Dziurawiec, & R. C. Annis (Eds.), *Explorations in cross-cultural psychology* (pp. 199-212). Lisse, the Netherlands : Swets.

SZALAY, L. B. (1981) Intercultural communication - a process model. *International Journal of Intercultural Relations*, 5, 133-146.

TANZER, N. K. (1995) Cross-cultural bias in Likert-type inventories : Perfect matching factor structures and still biased*? European Journal of Psychological Assessment*, 11, 194-201.

TANZER, N. K., GITTLER, G. & ELLIS, B. B. (1995) Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment*, 11, 170-183.

TANZER, N. K., GITTLER, G., & SIM, C. Q. E. (1994) A cross-cultural comparison of a Rasch calibrated spatial ability test between Austrian and Singaporean adolescents. In A. Bouvy, F. J. R. Van de Vijver, P. Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 96-110). Lisse, the Netherlands : Swets.

TANZER, N. K. & SIM, C. Q. E. (1991) *Test anxiety in primary school students : An empirical study in Singapore*. (Research Report 1991/6). Graz : Department of Psychology, University of Graz.

TANZER, N. K., & SIM, C. Q. E. (1992, July*) Personality of Singaporean adolescents. A replication study*. Poster presented at the 25th International Congress of Psychology, Brussels, Belgium.

TANZER, N. K., & SIM, C. Q. E. (1997) *Cross-cultural differences in the worry-achievement relationship : Evidence for construct bias or moderating influence of the educational system ?* Manuscript submitted for publication, University of Graz.

TANZER, N. K., SIM, C. Q. E., & MARSH, H. W. (1992) Test applications over cultures and languages : Theoretical considerations and empirical findings. *Bulletin of the International Test Commission*, 19, 151-171.

TANZER, N. K., SIM, C. Q. E., & SPIELBERGER, C. D. (1996) Experience and expression of anger in a Chinese society. The case of Singapore. In C. D. Spielberger & I. Sarason (Eds.), *Stress and emotion* (Vol. 16, pp. 51-65). Washington : Francis & Taylor.

TING TOOMEY, S., & KORZENNY, F. (Eds.). (1991*) Cross-cultural interpersonal communication*. Newbury Park, CA : Sage.

TOUBIANA, Y. (1994). Pictorial evaluation of test reactions. Petach-Tikva, Isreal : Peter.

TRIANDIS, H. C., & MARÍN G. (1983) Etic plus emic versus pseudoetic. A test of the basic assumption of contemporary cross-cultural psychology. *Journal of Cross-Cultural Psychology*, 14, 489-500.

VALENCIA, R. R., RANKIN, R. J., & LIVINGSTON, R. (1995) K-ABC content bias : Comparisons between Mexican American and White children. *Psychology in the Schools*, 32, 153-169.

VAN DE VIJVER, F. J. R. (1988) Systematizing item content in test design. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 291-307). New York : Plenum.

VAN DE VIJVER, F. J. R. (1991*) Inductive thinking across cultures : An empirical investigation*. Helmond : Wibro.

VAN DE VIJVER, F. J. R., DAAL, M., VAN ZONNEVELD, R. (1986) The trainability of abstract reasoning : A cross-cultural comparison. *International Journal of Psychology*, 21, 589-615.

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 J. W. Berry, Y. H. Poortinga, & J. Pandey (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., & POORTINGA, Y. H. (1997) Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13, 29-37.

VAN HAAFTEN, E. H., & VAN DE VIJVER, F. J. R. (1996) Psychological consequences of environmental degradation. *Journal of Health Psychology*, 1, 411-429.

VERSTER, J. M. (1983) The structure, organization, and correlates of cognitive speed and accuracy : A cross-cultural study using computerized tests. In S. H. Irvine & J. W. Berry (Eds.), *Human assessment and cultural factors* (pp. 275-292). New York : Plenum.

VODEGEL MATZEN, L. B. L., VAN DER MOLEN, M. W., & DUDINK, A. C. M. (1994) Error analysis of Raven test performance. *Personality and Individual Differences*, 16, 433-445.

WEHRLY, B. (1995) *Pathways to multicultural counseling competence : A developmental journey*. Pacific Grove, CA : Brooks/Cole.

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

WORD, C. O. (1977) Cross-cultural methods for survey research in Black urban areas. *Journal of Black Psychology*, 3, 72-87.

YANG, K. S., & BOND, M. H. (1990) Exploring implicit personality theories with indigenous or imported constructs, The Chinese case. *Journal of Personality and Social Psychology*, 58, 1087-1095.

YUI MIAO, E. S. C., & LIN, R.-F. (1990, July*) An exploratory study of the Anger Expression Scale : Comparing two groups of adolescents*. Paper presented at the 22nd International Congress of Applied Psychology, Kyoto, Japan.

Author's address :

Fons van de Vijver
Department of Psychology
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

European Review of Applied Psychology, 4th trimester, vol. 47, n° 4, pp. 280

Biais et équivalence dans l'évaluation inter-culturelle : une revue

Fons van de Vijver et Norbert K. Tanzer

(Version abrégée)

Dans une recherche inter-culturelle il est plus facile de poser la question de la signification psychologique des scores aux tests que d'y répondre. Pourtant il est d'une importance fondamentale de savoir si des notes à des tests obtenues dans différentes populations peuvent être interprétées de la même manière. Deux termes se sont trouvés associés à cette question, ceux de biais et d'équivalence. On dit qu'il existe un biais si des différences dans les notes au test ne correspondent pas à des différences semblables dans le domaine de la généralisation ; par exemple, supposons que les différences culturelles dans les notes d'intelligence reflètent des différences éducationnelles. Dans ces conditions, des différences de scores entre deux groupes nationaux devraient être interprétées en termes de facteurs éducationnels, alors que les différences individuelles à l'intérieur d'un pays pourraient dépendre de facteurs intellectuels. On présente une taxonomie tant du biais que de l'équivalence. Le biais peut être produit par le construct théorique (biais de construct), par la méthode, par exemple par la forme d'administration du test (biais de méthode) ou par le contenu de l'item (biais d'item). Un biais de construct survient lorsque le construct mesuré n'est pas identique dans les groupes culturels. Le terme "biais de méthode" a été créé parce qu'il dérive d'aspects décrits dans le paragraphe "Méthodologie" des publications rapportant des recherches empiriques. On peut envisager trois types de biais de méthode. Premièrement, le fait que les échantillons ne sont pas comparables sur des aspects autres que la variable cible peut amener à des biais de méthode (biais d'échantillon). Par exemple, des groupes culturellement différents diffèrent souvent en ce qui concerne l'arrière plan éducationnel et, lorsqu'on utilise des tests mentaux, ces différences peuvent perturber les différences réelles entre populations par rapport à la variable cible. Le biais de méthode se rapporte aussi aux problèmes dérivant des caractéristiques des instruments (biais d'instrument). Un exemple bien connu est celui de la familiarité avec le stimulus. Un dernier type de biais de méthode naît de problèmes d'administration (biais d'administration) comme il se produit par exemple quand il y a des problèmes de communication entre ceux qui dirigent l'entrevue et ceux qui répondent.

La biais d'item se rapporte à des distorsions se produisant au niveau de l'item. Les items biaisés ont une signification psychologique différente suivant les cultures, par exemple en raison de mauvaises traductions.

L'équivalence se rapporte au niveau de mesure auquel des scores peuvent être comparés dans plusieurs cultures. Trois niveaux d'équivalence sont possibles : le même construct est mesuré dans chaque groupe culturel, mais l'aspect fonctionnel de la relation entre les scores obtenus dans différents groupes est inconnu (équivalence structurelle) ; les scores utilisent la même unité de mesure mais une autre unité de mesure (équivalence d'unité de mesure), et les scores utilisent la même unité de mesure dans les populations, mais ont différentes origines (équivalence d'unité de mesure) ; les scores ont la même unité de mesure et la même origine dans toutes les populations (équivalence d'échelle complète). Les sources de biais les plus fréquemment rencontrés sont décrites ainsi que les moyens d'y remédier.

Sont décrits les éléments de base des études multilinguistiques. La traduction-rétrotraduction (c.a.d. une traduction dans le langage cible et une rétrotraduction indépendante dans le langage d'origine, suivies par une comparaison des deux versions dans le langage d'origine) et la méthode des comités (c.a.d. une traduction est préparée par un groupe de personne ayant souvent des expertises multiples sont mentionnées. En outre trois options dans une étude multilinguistique sont décrites. Premièrement, un instrument peut être traduit de manière littérale. Un aspect intéressant de cette option est qu'en principe une équivalence d'échelle complète existe ; toutefois une limitation importante est qu'il faut faire l'hypothèse d'une absence complète de biais de construct et de méthode. Deuxièmement, un instrument peut être adapté ; ceci consiste en une traduction littérale d'une partie de l'instrument et en une modification des stimulis probablement inadéquats dans le langage cible. Les adaptations sont moins sensibles au biais de construct et de méthode. Troisièmement, on peut construire un instrument complètement nouveau. L'avantage le plus important de cette construction est la possibilité d'adapter complètement l'instrument à la situation locale, son inconvénient majeur est l'impossibilité de faire aucune comparaison de notes entre groupes culturels.

Enfin, il est soutenu que le biais n'est pas une propriété intrinsèque d'un instrument mais une caractéristique des comparaisons inter-culturelles. Un item ou un test appropriés pour comparer les cultures A et B peuvent être inadéquats pour une comparaison entre A et C. En outre la présence de biais dépend aussi du but de l'étude. Si un chercheur est intéressé à savoir si un instrument mesure le même construct dans chacune des cultures examinées, la présence de certain biais peut poser moins de problèmes que s'il s'intéresse à comparer directement des groupes culturels par rapport à une variable cible. Par exemple, les différences inter-culturelles en ce qui concerne la familiarité vis à vis de stimulus ou la désirabilité sociale sont beaucoup plus faciles à traiter lorsque les scores ne sont pas directement comparés entre les cultures.