Translating tests

van de Vijver, F.J.R.; Hambleton, R.K.

Published in:
European Psychologist

Document version:
Peer reviewed version

Publication date:
1996

Translating Tests: Some Practical Guidelines

Fons Van de Vijver

Tilburg University, The Netherlands

and

Ronald K. Hambleton

University of Massachusetts at Amherst, USA

Key words: Test Translations, Test Adaptations, Bias, Guidelines

Abstract

With the increasing interest in cross-cultural research, there is a growing need for standard and validated practices for translating psychological instruments. Developing a psychologically acceptable instrument for another cultural group almost always requires more effort than a literal translation which all too often is the common practice. The adequacy of translations can be threatened by various sources of bias. Three types of bias are distinguished in this paper: (1) construct bias (related to non-equivalence of constructs across cultural groups), (2) method bias (resulting from instrument administration problems), and (3) item bias (often a result of inadequate translations such as incorrect word choice). Ways in which bias can affect the adequacy of instruments are illustrated and possible remedies are discussed.

Translating Tests: Some Practical Guidelines

Interest has grown steadily in cross-cultural comparisons over the last 20 years. For example, the 1995 Third International Mathematics and Science Study with over 40 participating countries and tests in over 30 languages is the largest cross-cultural comparative study of school achievement that has ever been conducted. The number of studies dealing with cross-cultural comparisons in PsycLit, an electronic journal publishing summaries of a wide variety of psychology journals, also reflects this increase (Van de Vijver & Lonner, 1995). To some extent, the increase is due to an increase in the number of journals with a masthead policy to publish mainly or exclusively cross-cultural studies such as the Hispanic Journal of Behavioral Sciences and Psychology and Developing Societies. However, the massive increase cannot be explained solely by the inauguration of new journals.

A considerably more important factor is the heightened interest in cross-cultural differences. Whereas in former days most of the cross-cultural research was carried out by psychologists who devoted most or all their research efforts to cross-cultural research, it is much more common today to find reports by researchers whose previous work took place within a single culture and who now are expanding their work to other cultures. An instrument that has shown good reliability and validity in one cultural context and has produced some interesting results, is applied elsewhere in order to examine cultural similarities and differences. For these researchers, a cross-cultural study does not mark the beginning of a research program in cross-cultural psychology. Instead, it is a natural extension of their previous work in a single culture.

The design, method, and analysis of cross-cultural studies have various unique features that are absent or less salient in intracultural studies. In this paper, the focus will be

on an important methodological aspect of cross-cultural research studies: translation of instruments. The application of an instrument in a new cultural group is more involved than simply producing text in another language, administering the translated instrument, and comparing the results (see, for example, Hambleton, 1993, 1994). There are many difficult questions to be addressed in multilingual studies: For example, does the construct apply to the target group or does it show an ethnocentric bias? Are the behaviors associated with the bias similar in the source and target groups? Is the measurement procedure (e.g., stimulus and response format) adequate for application in the target group?

A taxonomy of bias, meant here as a generic term for all kinds of factors that jeopardize the validity of intergroup comparisons, ranging from unobserved ethnocentrism in constructs to incorrect word choice in translations, will be presented in the first part of the paper. Three kinds of bias will be distinguished, depending on whether they are brought about by anomalies in the theoretical construct, instrument administration, or specific items. Depending on the kind of bias that can be expected, translations can amount to either the application of a literal translation of an instrument, of an adapted version, or of an entirely new instrument to measure the same construct. These three kinds of bias will be described in some detail in the second part of the paper. In the third and longest part of the paper, guidelines for test translations will be presented and described. Implications of these guidelines will be also be described.

Types of Bias in Test Translation

A distinction can be made among three types of bias in cross-cultural research (Van de Vijver & Poortinga, in press). The first is construct bias which is said to occur when the

construct that is measured by an instrument shows nonnegligible differences across cultures; both differences in conceptualization and in behaviors associated with the construct can underlie construct bias. A well known example of differences in conceptualization is provided by intelligence. In our measurements of intelligence there is an emphasis on cognitive performance such as reasoning (e.g., the Raven Progressive Matrices Tests) or previously acquired knowledge (in tests of crystallized intelligence). It has been shown repeatedly that everyday conceptualizations of intelligence are often broader and also include social aspects such as communication skills and even obedience (Serpell, 1993; Sternberg, 1985; Super, 1983).

An example of differences in behaviors is formed by the concept of filial piety; Ho (in press) found that behaviors associated with being a good son or daughter such as taking care of one's parents, conforming to their requests, and treating them well, are much broader in China than in most Western countries. Conclusions drawn on the basis of instruments that show construct bias, can be misleading when no reference is made to cross-cultural differences in the conceptualization or behaviors associated with the construct. Statements about differences in filial piety of Chinese and, say, German subjects, will be incorrect when they are based on an instrument that describes exclusively behaviors of these cultural groups.

It is important to examine the occurrence of construct bias by exploring any ethnocentric bias in our theory or operationalizations. Local surveys aimed at exploring the everyday conceptualizations of the construct and the behaviors associated with the construct provide an effective means to study construct bias. Such a survey could have various outcomes. It could support the validity of the existing instrument; it could also point to the inapplicability of particular items or sets of items (e.g., items about nuclear families in

communities that do not live in nuclear families). The findings of the survey are most consequential when there is a substantial lack of overlap of conceptualization or construct-characteristic behaviors. In such a case, substantial revisions of the conceptualization or instrument are required.

Construct bias is more likely to occur when an existing instrument is translated than when an instrument is simultaneously developed for different languages. In the latter case it is easier to avoid ethnocentric tendencies and to remove words and concepts in a source language that are not common in the two languages and cultures. A successful avoidance of ethnocentric tendencies in instruments may require a multicultural, multilingual team with an expertise in the construct under study.

Method bias is a generic term for validity-threatening factors that are related to instrument administration. Various sources of method bias are easy to imagine such as intergroup differences in social desirability; in response sets such as acquiescence; in familiarity with stimuli, response formats (e.g., multiple choice, Likert scales), or with testing situations in general; in physical conditions in which a test is administered; in subjects' motivation; in administrator effects; and in communication problems between the administrator and the persons taking the test. If present, method bias usually influences most or all items and hence, it will lead to differences in scores between groups that are to be attributed to the administration procedure and not to any intrinsic differences of the groups on the construct studied.

In order to examine method bias, an often neglected source of bias in cross-cultural studies, additional information has to be collected. An effective means is the application of additional methods to collect information about the same underlying trait such as is reflected

in the use of monotrait--multimethod matrices (e.g., Campbell & Fiske, 1959; Marsh & Byrne, 1993), also known as triangulation (e.g., Lipson & Meleis, 1989).

As an alternative, repeated test administrations can be applied. The procedure is particularly useful for mental tests. A study of the cross-cultural similarity of score changes from the first to the second test administration can give important clues about the validity of the measurement. When individuals from different groups with equal test scores on the first occasion have on average dissimilar scores on the second occasion, one can retrospectively doubt the validity of the first administration. Measurements of social desirability or studies of response sets can also address method bias (e.g., Fioravanti, Gough, & Frere, 1981; Hui & Triandis, 1989). Finally, method bias can be examined by administering the instrument in a nonstandard way, soliciting all kinds of responses from a respondent about the interpretation of instructions, items, response alternatives, and motivations for answers. Such a nonstandard administration provides an approximate check on the suitability of the instrument in the target group.

Item bias or differential item functioning (as it is sometimes called) is the last source of anomalies in instrument translations. It refers to instrument anomalies at the item level such as poor wording, inappropriateness of item content in a cultural group, and inaccurate translations. An item is biased if persons from different groups with the same score on the construct, commonly operationalized as the score on the instrument, do not have the same expected score on the item (Holland & Wainer, 1993; Shepard, Camilli, & Averill, 1981). For an unbiased item, knowledge of the total test score of a person does not contain information about group membership, while for a biased item it does. Various statistical techniques have been developed to detect item bias. The currently most popular technique for dichotomously-

scored items is the so-called Mantel-Haenszel procedure (Holland & Thayer, 1988; Holland & Wainer, 1993).

For test scores with interval-scale properties, other techniques can be applied. An example is an analysis of variance in which the item score is the dependent variable and culture and score level are the independent variables. The latter assumes that prior to the analysis the sample has been split in various score levels. A significant main effect of culture or of the interaction between culture and score level point to bias (more details can be found in Van de Vijver & Leung, in press).

Compared to construct and method bias, the examination of item bias is the least cumbersome. First of all, a large number of sophisticated statistical techniques to detect item bias are available; second, scrutinizing item bias does not require the collection of additional data as is the case for construct and method bias.

Options in Instrument Translations: Apply, Adapt, and Assemble

The nature and size of the bias that can be expected will have implications on the options available when translating an instrument. For instance, when a pilot study has shown the presence of method bias, various instrument alterations may be required in order to ensure the validity of the instrument in all groups. Depending on the changes that are required, instrument translators have three options: (a) to apply the instrument in a literal translation; (b) to adapt parts of the instrument; (c) to assemble an entirely new instrument (Van de Vijver & Leung, in press). Going from the first to the third option, there will be more changes required to make the instrument appropriate in the target group.

The translation options are related to the three types of bias distinguished before. The

first option, the application of the instrument, assumes that a literal translation of the instrument will yield an instrument in the target group that has good coverage of the theoretical construct and an adequate instrument format. In other words, both construct and method bias are then assumed to be absent and only item bias is examined. From a methodological perspective, the application option is straightforward. An instrument is translated into a target language or, in the case of a new instrument, simultaneously developed in two or more languages followed by an independent back-translation (Brislin, 1980; Werner & Campbell, 1970). After a comparison of the original and back-translated versions, possibly followed by suitable revision, the instrument is applied in the source and target cultures and the results are compared. The simplicity of the option probably explains its widespread usage.

However, the application option in multilingual studies may not address method and construct bias. When the instrument leaves important aspects of the construct in the target group unexplored, the application option will be inadequate and the instrument will have to be adjusted to the local context. Such an adjustment can take on various forms such as adaptations of the stimulus or response format or interviewer training, or the application of a multimethod approach.

The administration of not fully identical instruments in different cultural groups can complicate statistical analyses. Analyses of variance and t tests on total test scores assume identity of stimuli, and adjustments are required to deal with the stimulus dissimilarities. Some statistical techniques have scope for these dissimilarities. As an example, item response theory can be mentioned (see, for example, Hambleton, Swaminathan, & Rogers, 1991). Scores of examinees on the ability measured by the instrument (i.e., the latent trait) are independent of the particular stimuli that have been used to measure ability. Confirmatory

factor analysis also allows for incomplete overlap of stimuli.

Attractive as this may sound, the approach to overcome problems of partial overlap by applying sophisticated statistical techniques has limitations. When there is substantial overlap between the items administered in all groups, the approach will work well and the culture-specific items may well enhance the validity of the instrument in the local culture. However, when the overlap is small, the instrument will not have enough common material (referred to as the "anchor") on the basis of which scores can be compared across cultures and culture-specific items will add important aspects of the construct. A meaningful score comparison is then difficult to do.

The most severe instrument changes are usually required in the case of construct bias. Avoiding this type of bias may require the removal of particular items that are inappropriate in the new cultural context. For example, Van Haaften and Van de Vijver (in review) applied Amirkhan's (1990) Coping Strategy Indicator to Sahel dwellers. The item "watched more television than usual" had to be skipped because there was no electricity in the area of the study and television sets were uncommon.

The lack of overlap in conceptualization or in shared behaviors across cultures can become so small that an entirely new instrument has to be assembled. This is most likely to happen when an instrument that has to be developed in one cultural context, usually some Western country, contains various --implicit or explicit-- references to the local context of the test developer.

When entirely new instruments are assembled, the researcher is usually not interested in comparing average scores across cultures (e.g., in a t test) and there is more interest in the question as to whether the same psychological construct is measured in all groups. The

nomological network of the construct can then be compared across cultural groups using linear structural models (path models) or regression analysis.

Guidelines for Test Translations

In 1993 an international committee of psychologists was formed by the International Test Commission, consisting of members of various international organizations representing branches of psychology in which instrument translations play an important role. The committee has formulated a set of guidelines describing recommended practices in test translations. A preliminary report has been published (Hambleton, 1994); the final report will become available in 1996. The present section describes the 22 guidelines that were formulated; each guideline will be followed by a brief explanation.

The guidelines cover four domains: context (describing basic principles of multilingual studies), development (recommended practices in developing multilingual instruments), administration (issues in instrument administrations), and documentation/score interpretation (related to interpretation and cross-cultural comparisons of scores).

The context guidelines are as follows:

1. Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.

The guideline expresses a basic principle of cross-cultural research: avoid construct, method, and item bias as much as possible. Multilingual studies should be geared towards generating interpretable patterns of intergroup similarities and differences. Without sufficient precautions against alternative interpretations, intergroup differences tend to be multi-

interpretable (Poortinga & Malpass, 1986). For example, differences in observed scores on the Raven's Progressive Matrices Test obtained in two widely different cultural groups could be due to valid intergroup differences in intelligence, but also to intergroup differences in familiarity with the instrument or the testing situation, in educational background, in motivation, etc. When no information is available to rule out alternative interpretations, it will become difficult to interpret observed differences.

The guideline does not state that bias sources should be eliminated but adopts the more realistic position that they should be minimized. When cultural and linguistic differences between the groups studied are not too big, it may be realistic to pursue the elimination of bias; however, when large cultural distances have to be bridged by the instrument, particular sources of bias may be impossible to overcome. For example, when the Raven's test of intelligence has been administered to literate and illiterate groups, it will be unrealistic to assume equality of the groups on factors related to method bias such as familiarity with stimuli or with the testing situation in general. Measures of bias-related factors such as a measure of stimulus familiarity or previous test exposure can often be added to corroborate a particular interpretation of intergroup differences. These measures can be introduced as covariates in an analysis of covariance in order to examine as to whether there are remaining intergroup differences after statistical correction for these biasing factors.

2. The amount of overlap in the constructs in the populations of interest should be assessed.

The guideline refers to construct bias and stresses the need to assess instead of assume similarity of meaning and of construct-characteristic behaviors across cultural groups. Pilot studies aimed at identifying construct-characteristic behaviors and cooperation with local

experts are tools to examine construct bias.

The guideline expresses a principle recurring in several others: the validity of an instrument in multilingual studies cannot be taken for granted but has to be demonstrated.

The guidelines on instrument development are as follows:

3. Instrument developers/publishers should insure that the translation/adaptation process takes full account of linguistic and cultural differences among the populations for whom the translated/adapted versions of the instrument are intended.

The translation of a test requires a thorough knowledge of both the target language and the culture. Hambleton (1994, p. 235) provides a useful example to illustrate the point. In a Swedish-English comparison of educational achievement the following item was administered:

> Where is a bird with webbed feet most likely to live?
>
> a. in the mountains
>
> b. in the woods
>
> c. in the sea
>
> d. in the desert.

The Swedish translation rendered "webbed feet" as "swimming feet," thereby providing a cue about the correct answer. Such language- or culture-specific elements can easily slip into translations (or remain unnoticed when they are present in the original instrument). In back-translation procedures, aiming at a verbatim comparison of original and back-translated versions, these problems may remain undetected.

4. Instrument developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are

appropriate for all cultural and language populations for whom the instrument is intended.

The terms and concepts used in the instrument should be appropriate to all cultural groups involved. It is important to indicate which measures have been taken to ensure the translatability of instruments and to reduce the problem of miscommunication. Various rules have been formulated as to how translatable instruments can be designed. For example, Brislin (1986) has formulated various guidelines to optimize the translatability of an instrument, the most important of which are given here (p. 143-150):

- Use short and simple sentences and avoid unnecessary words (unless redundancy is deliberately sought).
- Employ the active rather than the passive voice because the latter is easier to comprehend.
- Repeat nouns instead of using pronouns because the latter may have vague referents; thus, the English "you" can refer to a single or to a group of persons.
- Avoid metaphors and colloquialisms. In many cases their translations will not be equally concise, familiar, and captivating.
- Avoid verbs and prepositions telling “where” and “when” that do not have a precise meaning, such as “soon” and “often.”
- Avoid possessive forms where possible because it may be difficult to determine the ownership. The ownership such as “his” in “his dog” has to be derived from the context of the sentence and languages vary in their system of reference.
- Use specific rather than general terms. Who is included in “members of your

family" strongly differs across cultures; more precise terms are less likely to run into this problem.

5. Instrument developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations.

Various formal instrument characteristics such as its response format can jeopardize the validity of cross-cultural comparisons. For instance, the ability to solve items in a multiple-choice format requires previous knowledge and experience. Thus, alternatives often show subtle differences in meaning. Another example is the ability to deal with speed tests that often requires a delicate balance between speed and accuracy. The application of instruments among groups without relevant testing experience can be troubled by such unexpected intergroup differences. When there is a real danger that such factors will affect performance, test developers may want to provide information about how they have dealt with the problem (e.g., lengthy test instructions or a repeated test administration).

6. Instrument developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.

The guideline is related to the previous two, stressing the importance of examining the familiarity of stimulus features. The guideline is often addressed in mental testing; the notion that cognitive tests to be utilized in various cultural groups should be "culture-free" (Cattell, 1940), "culture-fair" (Cattell & Cattell, 1963), or "culture-reduced" (Jensen, 1980) originates in the recognition of the importance of stimulus familiarity. Stimulus familiarity is often difficult to measure. Methods to study method bias such as repeated test administrations and multimethod approaches can be applied to evaluate intergroup differences in stimulus

familiarity. Cross-cultural differences that are not invariant across repeated test administrations or different methods to measure the same construct point to differential stimulus familiarity.

Items with a different ecological validity in the cultural contexts in which they will be applied, are not suitable for cross-cultural comparison. If there is reason to suspect differences in ecological validity, a pilot study can be carried out addressing the issue.

The concept of stimulus familiarity has been most often discussed in the area of mental testing. However, the concept also refers to other psychological constructs. Cross-cultural comparisons of scores on personality questionnaires that contain items with a low ecological validity will have dubious validity.

7. Instrument developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the translation/adaptation process and compile evidence on the equivalence of all language versions.

The judgmental evidence described in the guideline involves the application of standardized translation procedures, such as translation--back-translation. Similarity of the original and back-translated versions are taken to indicate appropriate translation. The procedure is particularly useful when the researcher does not know the target language because it gives an evaluation of the quality of the target language version that is accessible to the researcher. At the same time, an adequate back-translation does not guarantee an appropriate target language version. For example, the procedure favors literal translations while readability and naturalness of the target language version is often hardly checked; literal translations can produce stilted language, a feature that may not be detected by a back-

translation. Translators who know that their work will be back-translated may favor such literal translations.

The quality of translations of texts that are difficult to translate may benefit from an approach in which not a single bilingual but a whole group of persons participate in the translation process. Particularly when such a group combines linguistic and psychological expertise, the quality of the translation may be superior than what is found using a translation--back-translation approach with two translators--one doing the source to target language translation, and the other translator doing the reverse translation (see, Hambleton, 1993).

8. Instrument developers/publishers should insure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the instrument.

The design of the study in which source and target language versions are compared should allow for a rigorous test of equivalence. Various designs have been proposed (cf. Hambleton, 1994, p. 237-238). For example, in one popular design, bilinguals take source and target versions of the test. An obvious problem of the design is the need to find a sufficiently large sample of bilinguals. Furthermore, bilinguals may constitute an atypical sample of the population because they are usually better educated. A major asset of this design is control on the similarity of the sample taking both versions.

This factor is not controlled in the most common design to study equivalence, a design in which source-language monolinguals take the source-language version and target language monolinguals take the target-language version. The design confounds population and translation characteristics; an intergroup difference on a particular item can be attributed to poor translation (e.g., inadequate word choice) and/or to population characteristics (e.g.,

the item is more attractive in group A than in group B, or persons in group A are simply more capable than persons in group B).

9. Instrument developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the instrument, and (2) identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations.

An evaluation of the appropriateness of the translation should not only be based on judgmental evidence such as provided in a translation--back-translation procedure but also on statistical evidence. Empirical data should be collected and properly analyzed in order to examine the equivalence of the source and target versions of an instrument. Various techniques can be used for that purpose (see, Hambleton, 1993; Van de Vijver & Leung, in press). Frequently applied is factor analysis, either exploratory (e.g., Barrett, 1986) or confirmatory (e.g., Watkins, 1989).

10. Instrument developers/publishers should provide information on the evaluation of validity in all target populations for whom the translated/adapted versions are intended.

Transfer of validity (e.g., construct and predictive validity) from one cultural context to the other cannot be taken for granted but has to be demonstrated. Instruments that have good validity in one cultural group may lose some of their psychometric properties after translation. Larger cultural distances will generally jeopardize the validity more.

11. Instrument developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations.

The guideline stipulates the need for item bias analyses, scrutinizing the equivalence

on an item by item basis (e.g., Holland & Wainer, 1993). Various steps are possible when item bias is found. First, the items can be taken to constitute a threat to the validity of the instrument and can be eliminated. When the bias of all items has been examined, cross-cultural comparison can be restricted to the presumably unbiased set. Second, item bias can be seen as pointing to interesting cross-cultural differences that require further examination and explanation. For example, commonalities among the biased items can be sought. Unfortunately, such commonalities are often hard to find (e.g., Scheuneman, 1987). A third possible step is described in the next guideline.

> 12. Nonequivalent questions between versions intended for different populations should **NOT** be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content validity of scores reported for each population separately (emphasis in original).

Nonequivalent questions will be invalid in cross-cultural comparisons of scores (unless, as indicated before, a statistical technique is applied that can handle stimulus dissimilarities such as item response theory and linear structural modeling); however, they may be adequate in intracultural use of the instrument adding to its reliability and validity.

The administration guidelines are as follows:

> 13. Instrument developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions.

Administration problems can often be detected in small pilot studies in which the instrument is applied in a nonstandard way soliciting various responses from the respondents. Careful observation and asking respondents to paraphrase items and to provide reasons for

their responses will help to identify such problems.

14. Instrument administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.

A literal translation of stimuli is often the preferred choice in multilingual studies. However, test administrators should be aware of the specific problems that this may create; for example, particular examples may not be very obvious to some groups; the test instructions may contain some implicit information that may not be clear to individuals from different cultural groups. As another example, Raven's test of intelligence has two series of items, one that can be solved using more perceptual strategies while the second series is more difficult and requires analytical strategies. The test instructions contain only item examples that use a perceptual strategy. The change of item content may be confusing to individuals who have little or no test experience. In general, the application of identical stimuli or literally translated stimuli does not guarantee that the instrument is appropriate in each cultural group; a close examination of validity moderating factors is vital in all multilingual studies.

15. Those aspects of the environment that influence the administration of an instrument should be made as similar as possible across populations for whom the instrument is intended.

Environmental conditions of laboratory testing may be easy to replicate elsewhere, but physical conditions of field research tend to be idiosyncratic and hard to replicate elsewhere. It is therefore important that test administrators are made cognizant of the main environmental variables that should be kept in mind. For example, in an administration of computerized speed tests, body posture, distance to the screen, and intensity of ambient light

are among the factors that have to be considered.

16. Instrument administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.

When an instrument will be applied in a new cultural context, instrument developers need to know sources of unwanted intergroup differences. Pilot studies can help to detect these differences. The test instructions are an important aid in the reduction of the differences. Lengthy test instructions, containing various examples and exercises, can go a long way to minimize these differences.

17. The instrument manual should specify all aspects of the instrument and its administration that require scrutiny in the application of the instrument in a new cultural context.

Test developers will have gained relevant information about the specific issues that arose in the test translation process. Administrators of the test can benefit from this experience when the manual gives all the necessary details. The test manual should describe potential problems in order to avoid their repetition.

18. The administration should be unobtrusive and the administrator--examinee interaction should be minimized. Explicit rules that are described in the manual for the instrument should be followed.

An important source of errors in cross-cultural comparisons can be the uncontrolled aspects of administrator-examinee interactions, particular in nonstandardized testing situations such as unstructured interviews. The manual should specify standard problems and their solutions. For example, the manual for an intelligence test should specify the correctness

and incorrectness of answers and, if applicable, the supplementary questions to be asked following partially correct answers. When such rules are not specified in the manual, it will be impossible to standardize test scoring.

The guidelines on documentation/score interpretations are as follows:

19. When an instrument is translated/adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence.

In the previous section, distinctions were made among applying, adapting, and assembling tests in translations. When tests are applied, the test content is not changed and the new test is a direct translation of the test in the source language. When tests are adapted or new tests are assembled, test users should be informed of all changes introduced to enhance the validity in a new cultural context. Furthermore, the equivalence of the source and target language versions of the test should be documented. Linguistic and statistical evidence such as a specification of the translation procedure, the results of an item bias analysis or of a factor analysis comparing the loadings across cultural groups, should be described (Van de Vijver & Leung, in press). Without such evidence, it will be difficult for potential test users to determine the adequacy of the test in the new context.

20. Score differences among samples of populations administered the instrument should **NOT** be taken at face value. The researcher has the responsibility to substantiate the differences with other empirical evidence (emphasis in original).

Observed intergroup score differences can often be interpreted in several ways (Poortinga & Malpass, 1986). If a researcher embraces particular interpretations, he/she should provide evidence to confirm these interpretations or to disprove alternative

interpretations. In order to provide the evidence, additional measures will often be required; for example, disproving that a cultural difference is due to differential social desirability, acquiescence, or stimulus familiarity, will often require measurement of these factors. The approach not to take observed score differences at face value may seem unduly restrictive; however, upon closer examination it is a natural consequence of the poor controls that are available in cross-cultural comparative studies. In cross-cultural comparisons, groups that differ on many dimensions are often used; in many cases we are only interested in a single or a few of these but it would be naive to act as if the other differences do not exist. Therefore, in cross-cultural psychology we need to safeguard our data against alternative interpretations.

21. Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.

The guideline refers to an important concept in cross-cultural comparisons: the comparison scale (e.g., Van de Vijver & Poortinga, in press). When an instrument has been applied in two cultures, the measurement level of three scales has to be observed: the first two are the measurement level of the scale in the two groups, the third one is the measurement levels of the score comparisons.

Suppose that an anxiety measure using Likert response scales (i.e., strongly disagree, disagree, neutral, agree, strongly agree) has been administered in two cultural groups. Let us assume that the sum of the item scores defines an interval scale in the two cultural groups. What is the measurement level at which the scores can be compared across cultures? Are individual differences within a single group measured at the same measurement level as individual differences across groups? An answer to the question depends on the presence of bias. When no bias occurs, individual differences within and across groups are measured at

the same level. However, bias will tend to lower the measurement level of the comparison scale.

Suppose that a few items have been poorly translated or are inappropriate in the second group. The biased items will constitute an offset in the comparison scale. When these items are not removed, the measurement level of the comparison scale may be ordinal. When the anxiety measure suffers from construct or method bias, the measurement level of the comparison scale can become even lower.

According to the present guideline, cross-cultural score comparisons can only be made at the level of the comparison scale that has been established. The comparison of cross-cultural differences in a t test or analysis of variance assumes the absence of any kind of bias; when the equivalence of the instrument across cultural groups has not been examined, the conclusions of such an analysis are often open to alternative interpretations.

22. The instrument developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance on the instrument, and should suggest procedures to account for these effects in the interpretation of results.

The test manual should specify all relevant examinee/respondent and context variables that have been examined in the development of an instrument such as relevant cultural characteristics of the target groups, socio-economic status, age, gender, and education. When the results of these analyses are presented in the manual, test users will know which factors will be relevant in their use of the instrument and how to account for these factors.

Implications

Translating psychological instruments for use in other cultural and linguistic groups is more involved than simply translating text into another language. Various sources of bias can threaten the adequacy of translations. Distinctions were made in this paper among three types of bias, depending on whether the bias resides in the construct or its characteristic behaviors (construct bias), in the measurement procedure (method bias), or in the separate items (item bias). Simple translation--back-translation procedures are meaningful only when construct and method bias do not play a role. When these play a role, more instrument adaptations will be required. Hambleton (1994) stressed the need to demonstrate the similarity of meaning of the directions, items, and even the scoring guides for the instrument in all linguistic groups involved.

The presence of bias depends on various factors. For example, the likelihood of bias will increase with the cultural distance between the groups involved. For example, comparisons of groups with strongly dissimilar backgrounds can easily suffer from method bias (e.g., differential stimulus familiarity or social desirability). The presence of bias will also depend on the nature of the construct. Broadly defined constructs described by heterogeneous behaviors will be more liable to construct bias.

A “cookbook” specifying the types of bias which can arise in practice and when they can be expected is impossible to give. Fortunately, such a cookbook is hardly ever required. Awareness that bias can play a role is an important step towards its detection and resolution. A combination of awareness and linguistic and psychological expertise will often suffice to yield high quality translations.

The advent of cross-cultural research described in the introductory section of the paper

creates a need for standard practices to carry out intergroup comparisons. The guidelines described in the paper are an attempt to formalize recommended practice in test translations. Hopefully, such guidelines will become standard practice in cross-cultural research and will help to minimize the impact of bias on cross-cultural measurement.

References

Amirkhan, J. H. (1990). A factor-analytically derived measure of coping: The Coping Strategy Indicator. Journal of Personality and Social Psychology, 59, 1066-1074.

Barrett, P. (1986). Factor comparison: An examination of three methods. Personality and Individual Differences, 7, 327-340.

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.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait--multimethod matrix. Psychological Bulletin, 56, 81-105.

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

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

Fioravanti, M., Gough, H. G., & Frere, L. J. (1981). English, French, and Italian adjective check lists: A social desirability analysis. Journal of Cross-Cultural Psychology, 12, 461-472.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests:

A progress report. European Journal of Psychological Assessment, 10, 229-244.

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

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

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

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

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

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

Lipson, J. G., & Meleis, A. I. (1989). Methodological issues in research with immigrants. Special Issue: Cross-cultural nursing: Anthropological approaches to nursing research. Medical Anthropology, 12, 103-115.

Marsh, H. W., & Byrne, B. M. (1993). Confirmatory factor analysis of multigroup--multimethod self-concept data: Between-group and within-group invariance constraints. Multivariate Behavioral Research, 28, 313-349.

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.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test

items. Journal of Educational Measurement, 24, 97-118.

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

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

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

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

Van Haaften, E. H., & Van de Vijver, F. J. R. (in review). Psychological consequences of environmental degradation.

Van de Vijver, F. J. R., & Leung, K. (in press). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), Handbook of Cross-Cultural Psychology. Boston: Allyn & Bacon.

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

Van de Vijver, F. J. R., & Poortinga, Y. H. (in press). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment.

Watkins, D. (1989). The role of confirmatory factor analysis in cross-cultural research. International Journal of Psychology, 24, 685-701.

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.