DOCUMENT RESUME

ED 337 481 TM 017 296

AUTHOR Hambleton, Ronald K.; Bollwark, John
TITLE Adapting Tests for Use in Different Cultures: Technical Issues and Methods.
PUB DATE 91
NOTE 44p.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cultural Differences; Educational Assessment; English; Foreign Countries; Guidelines; Spanish; Test Format; *Testing Problems; Test Validity; *Translation
IDENTIFIERS Scholastic Aptitude Test; *Test Adaptations

ABSTRACT

The validity of results from international assessments depends on the correctness of the test translations. If the tests presented in one language are more or less difficult because of the manner in which they are translated, the validity of any interpretation of the results can be questioned. Many test translation methods exist in the literature, but most are rather limited in their appropriateness. This paper reviews the issues and methods associated with test translations or adaptations, and presents some new results based on applications of item response theory (IRT) to establishing test guidelines. Guidelines are offered for establishing test equivalence based on a review of past studies and current methods, particularly methods that involve double test translations and IRT methods. An example of translation equivalence is drawn from the study by W. H. Angoff and L. L. Cook (1988) on the equating of English and Spanish versions of the Scholastic Aptitude Test. Two figures illustrate the discussion. A 33-item list of references is included. (SLD)

ADAPTING TESTS FOR USE IN DIFFERENT CULTURES: TECHNICAL ISSUES AND METHODS[1]

Ronald K. Hambleton, & John Bollwark
University of Massachusetts at Amherst

Abstract

In recent years, there has been considerable interest in international assessments of educational achievement. The validity of the results from these international assessments (such as the one recently completed in several countries in the areas of mathematics and science) depends on the correctness of the test translations. If the tests presented in one language are more or less difficult because of the manner in which the tests are translated, the validity of any interpretations of the results can be questioned. Many test translation methods currently exist in the literature, but most are rather limited in their appropriateness. In addition, the problem itself is one of considerable difficulty.

The purposes of this paper are to review the issues and methods associated with test translations or adaptations, to present some new results based on applications of item response theory (IRT) to establishing test equivalence, and to offer a set of guidelines for conducting test translation studies based upon a review of past studies and current promising methods, especially methods involving double test translations and IRT methods.

[1]To appear in the Bulletin of the International Test Commission, 1991.

Adapting tests for use in populations other than those the tests were designed for has its roots in the beginnings of intelligence testing. Psychologists around the world readily saw the potential of Binet's intelligence test for diagnostic and selection purposes, and adapted it for use in various populations of interest. In those first test adaptations, the process usually was a direct translation of the test.

More recently, adapting tests for use in populations other than those for whom the test was designed has been fueled by an interest in providing a basis for cross-population comparisons. Researchers interested in quantifying differences in intelligence and other traits in different populations must rely on test adaptations. Also, in countries such as the United States, issues of test bias have initiated an interest in adapting tests so that they are more relevant and thus "fair" to specific segments of a particular population. The adaptation process in these cases should ideally consist of translating a test from one language to another, with consideration given to the linguistic and cultural relevance of the translated version and to the "equivalence" of the different versions of the test.

Validly translating a test from one language to another and establishing the equivalence of the original and translated versions is a complex process. It is important that the process be better understood since test translations will play an increasingly important role in future testing activities. The main reason for this is that we are increasingly viewing our world from a multicultural perspective and therefore there is a need to (1) understand the similarities and differences that exist between populations and (2) provide unbiased

testing opportunities across different segments of a single population. Testing across populations provides a means for accomplishing these goals.

For example, in 1988, the International Assessment of Educational Progress (IAEP) was implemented (Lapointe, Mead, & Phillips, 1989). The goal of this project was to assess achievement in a common core of science and mathematics for 13-year-olds in five countries and four Canadian provinces. In order to accomplish this goal, test items in English were translated into several different languages. Also administered were questionnaires regarding students' school experiences and attitudes towards mathematics and science.

This expensive and time-consuming assessment project was undertaken because the results provided potential insights into what aspects of different populations influence the attainment of successful educational goals. One result from this study was that students from the United States scored lowest in mathematics achievement while Korean students scored highest. What reason or reasons are responsible for these differences? An answer to this question may be of substantial use in improving mathematics education in the United States and therefore is of vital importance to our society. Without cross-cultural assessment projects such as the IAEP, answers to these types of questions cannot be obtained. Without a proven methodology for evaluating the equivalence of the original and translated assessment instruments, a valid basis for these types of comparisons remains in question.

The purposes of this paper are to provide an overview of language translation of tests and inventories, and the methods used to establish translation equivalence. The discussion that follows focuses on tests

with the understanding that much of the discussion is generalized to occupational and interest inventories as well. The following topics are discussed: (1) The Purposes of Test Translations, (2) Past and Present Trends of Test Translation Use, (3) Problems Associated with Translating Tests, (4) Methods of Establishing Translation Equivalence, (5) Review and Selection of Methods, (6) Item Response Models in Establishing Translation Equivalence, and (7) Example of a Translation Equivalence Study.

The Purposes of Test Translations

Developing a test for use in a specific population can be accomplished by either (1) developing the test within the cultural boundaries of the population of interest or (2) translating an existing test so that it is appropriate for the population of interest. If the purpose of developing a population-specific test is to reduce cultural bias in the test scores, either one of the development methods may be used; however, certain purposes require the use of the second method - test translation.

The first purpose that requires the use of test translation is the economical development of tests that are valid for use in specific populations or sub-populations. Some nations do not have qualified personnel available for test development and validation. In such cases, translating existing tests is the only viable alternative for test development.

A second purpose that requires the use of test translation is providing a basis for comparisons between populations (either distinct populations or within a population whose members' primary language or other cultural traits differ). A recent example is the 1988

International Assessment of Educational Progress (IAEP). This assessment project required translating science and mathematics test items from English to French, Korean, and Spanish in order to make comparisons of achievement in these subjects across several populations (Lapointe, Mead, & Phillips, 1989).

While both purposes for test translations are valid, it is the second purpose - cross-population comparisons - that are of particular interest since test translations are the only alternative for allowing such comparisons. Nations lacking qualified personnel for test development may have the option of acquiring such expertise, thus reducing the need for test translations; however, those involved in cross-population comparisons are more dependent on the use of translation techniques.

Past and Present Trends of Test Translation Use

The first test translated into another language was the Binet-Simon intelligence test. Henry Goddard translated the test from French to English in 1911 for use at the Vineland Training School for the mentally retarded in New Jersey. By 1916, the Binet-Simon test had been translated into seven languages (Stanley & Hopkins, 1972).

Since these early test translations, numerous tests have been translated into the primary language of the examinees to be tested. Some examples include the Otis Group Intelligence Scale, Wechsler Intelligence Scale for Children, and the Wechsler Adult Intelligence Scale. However, criticism of test translations has also paralleled the use of this technique. Underlying much of the criticism were problems in (a) establishing equivalence in vocabulary, (b) determining the

dominant language of target population examinees, and (c) cultural differences in responding to stimuli.

Despite these criticisms, tests (and questionnaires/inventories) are continually being translated for use in target populations. The reasons for this are clear. First, the development of population-specific tests for certain purposes requires the use of test translations. Second, empirical studies support the use of test translations. Partial or total equivalence of translations have been reported by, for example, Hulin, Drasgow, and Komocar (1982); Hansen and Fouad (1984); Hulin and Mayer (1986); Fouad and Hansen (1987); and Candell and Hulin (1986). For these two reasons, test translations have become an important aspect of test development work, particularly in the areas of intelligence and aptitude tests.

Problems Associated with Translating Tests

The use of tests in populations other than those the test was designed for has raised concerns since the beginnings of intelligence testing (Samuda, 1983). In the case of test translations, it is assumed that enough differences between the populations of interest exist to warrant the development of a translated version of a test - it is identifying these differences and incorporating solutions to minimizing them that underlie many of the problems associated with translating tests. Four problems,which will be considered next, are especially important.

Identifying and Minimizing Cultural Differences

An initial problem in the translation process is identifying the cultural differences between the source and target populations that may affect examinee test performance. Among these cultural traits are

examinee motivation, values, experiences, and degree of test anxiety (van de Vijver & Poortinga, 1991). Cross-cultural researchers have provided numerous examples of how these cultural variables can influence the testing process. For example, van de Vijver & Poortinga (1991) point out difficulties experienced by Porteus in the administration of the Porteus Maze Test:

> . . . Porteus . . . for instance, found it difficult to persuade Australian aboriginal subjects to solve the items by their own effort rather than in cooperation with the tester. As another example, it can be mentioned that the Maze Test, which is a paper-and-pencil test, has been applied among groups from which the members had never touched a pencil before.

This example, and others, though they do not deal directly with test translations, points out that cultural differences between the source and target populations can affect examinee performance. It is therefore important to identify these cultural differences as a first step towards minimizing these effects. A further complication is that cultural differences must be considered for all parts of the testing process including test instructions, test items (content, response format, response mode, and symbol usage), administrator-examinee interactions and testing environment (Berry & Lopez, 1977; van de Vijver & Poortinga, 1991).

Identifying the Appropriate Language for Testing Target Population Examinees

Problems associated with identifying the appropriate language to be used when testing examinees in the target population sometimes arise. Problems may arise because of varied dialects within the target language (Berry & Lopez, 1977; Olmedo, 1981). Olmedo (1981) noted: ". . . it is not uncommon to find that many tests written in formal Spanish are used

inappropriately with populations that speak substantially different Spanish dialects." Unless examinees are being tested on their abilities with a formal language, at a minimum, even if translations to accommodate varied dialects are not being done, it is important to identify the dialects spoken in the target language (and what members of the target population speak them) in order to make valid test score interpretations.

An even more complex problem associated with language and test translations is determining the most appropriate language for testing bilingual target examinees. DeAvila and Havassy (1974) pointed out that, because a person speaks a language, it can not be assumed that s/he can read and therefore be non-verbally tested in that language (neither can it be assumed that a person thinks in that language). Moreover, a person may only be a functionally receptive bilingual. For example, "children from homes where parents prefer to speak Spanish may themselves be only functionally receptive bilinguals. They may understand Spanish but express themselves in English. The situation with the parents may be the reverse" (Olmedo, 1981). These situations point out the importance of understanding the extent of bilingualism and its implications for testing in bilingual target examinees. Failure to determine the most appropriate language for testing the target population can seriously undermine the validity of translating a test from the start.

Finding Equivalent Words or Phrases

A third problem associated with language and test translations is finding, if they exist, words or phrases that are equivalent in the source and target languages. For example, in a Spanish translation of

the Strong-Campbell Interest Inventory (II), Hansen and Fouad (1984) had difficulties finding an equivalent Spanish translation for the English word "argument" (the authors report similar difficulties with seven additional items).

In an attempt to alleviate the problem of non-equivalent words or phrases in the source and target languages, a process known as decentering is sometimes used. Decentering refers to the modifying of words or phrases in either initially the source version of a test or later, in both language versions of a test in order to achieve item equivalence. For example, the Spanish word "paloma" is equivalent to either "dove" or "pigeon" in English (Swanson & Watson, 1982) and therefore a test item in English that requires making a distinction between a dove and a pigeon would be difficult to translate into Spanish. The original item in English could be decentered by using a pair of terms that have similar meanings within the context of the item, and have equivalent terms in Spanish, thus allowing for a translation of the item.

Hulin and Mayer (1986) pointed out, however, that decentering may introduce psychometric nonequivalence between the original and translated item:

> Decentering produces translated material with smooth and natural terms in both versions. The price paid for such linguistic achievement may be that neither version is centered in either culture or language. Decentering should produce symmetrical translations with equal degrees of familiarity, colloquialism, and idiosyncrasy in both languages but fidelity to neither. The optimally decentered version, chosen through a mixture of back translations and discussions among translators, may introduce serious questions about psychometric equivalence between the two versions. For instance, an English version of a questionnaire that contained the phrase "Once in a blue moon" (to describe the frequency of promotions) might result in a decentered Spanish phrase, "Every time a bishop dies."

> Linguistically and ethnographically, the two versions are equivalent. The price of linguistic smoothness, however, may be paid in the coin of psychometric nonequivalence.

Unfortunately, it is difficult to get a sense of the extent and appropriateness of decentering used in specific test translations from the literature; descriptions of test translations often report only whether decentering was used or not. Useful information for evaluating the decentering process might include the percentage of items decentered and illustrative examples of how the decentering was accomplished.

Finding Competent Translators

Lastly, there are also practical problems associated with test translations. Translators familiar with the source and target language and competent in the material covered by the source test can be difficult to find. The problem of finding competent translators becomes compounded when the test covers a specialized content domain (for example, medicine).

Summary

Four problems associated with translating tests have been discussed. The extent to which each of the four points is a problem in translating a test will, of course, vary depending on the characteristics of the test and of the source and target populations. For example, it may be more difficult to identify and minimize cultural differences for a test with a high degree of verbal loading than a test that makes greater use of symbols. Moreover, if the characteristics of the source and target populations differ greatly, identifying and minimizing cultural differences will be more difficult than for source and target populations with similar or overlapping characteristics. Translating a test from one language to another and maintaining its

validity with respect to a specific purpose can be an exceedingly complex process. Being aware of the many potential problems in translating tests may help to minimize the errors associated with the translation process.

Methods of Establishing Translation Equivalence

Equivalence of test items is defined as the direct comparability of test items and the scores derived from them in terms of psychometric meaning. Thus, test items are equivalent if they measure the same behaviors across the populations of interest and examinees with equal amounts of ability within the populations have equal probabilities (within the limits of measurement error) of answering the items correctly.

A review of the literature on test and inventory translations indicated that many different methods have been used to establish the equivalence between source and translated instruments. Some of the methods are more commonly used than others; however, a comprehensive review of most or all of the available methods seemed useful. These methods include those that are used both before and after examinee responses have been collected. Each of the methods will be discussed mostly in terms of tests and test items with the understanding that these discussions generally apply to questionnaires and inventories as well.

The methods of establishing equivalence between original and translated test items can be viewed as an extension of the methods used for identifying item bias. In bias studies, the focus is on the items or scores derived from them for a single test. Establishing translation equivalence extends this focus to the items or scores derived from them

on two tests - the original test and either the initial translation or the back translated version of the original test. The presence of more than one version of a test on which to compare scores gives rise to the various methods of establishing translation equivalence to be discussed.

There is also a similarity in the methods used to establish translation equivalence and to identify biased items. In each case, both (a) judgmental and (b) statistical methods may be used. Judgmental methods of establishing translation equivalence are based on a decision by an individual or a group on the degree of each item's translation equivalence. In contrast, statistical methods establish translation equivalence based on the analysis of examinee responses to some combination of the original, translated, or back translated test items. The use of judgmental and statistical methods is not necessarily independent. Judgmental methods are often used as preliminary checks of translation equivalence before the tests are administered and statistical methods applied to the test scores.

The classification scheme adopted for identifying methods of establishing translation equivalence in this paper is based on whether judgmental or statistical methods are used. In addition, it is also useful to identify whether a single or back translation is used. Therefore, four categories of methods can be identified:

1.A Judgmental single-translation methods

1.B Judgmental back-translation methods

2.A Statistical single-translation methods

2.B Statistical back-translation methods

Figure 1 provides an overview of the current methods within each of the categories. These seven methods are considered next.

- -

Insert Figure 1 about here.

- -

Judgmental Methods

As stated previously, judgmental methods of establishing translation equivalence are based on a decision by an individual or a group on the degree of each item's translation equivalence. Thus, judgmental methods provide a subjective viewpoint on the question of equivalence.

1.A.1 *Post-translation probes*. In this method, one or more samples of target examinees answer the translated version of an item and are then asked about the meaning of their answers. Evidence of translation equivalence is obtained if the responses given by a high percentage of the examinees questioned reflect a reasonable interpretation of an item in terms of cultural and linguistic understanding. The main judgmental aspect of this method is deciding what responses by target examinees about the meaning of their answer to an item are considered reasonable.

The use of this method can provide valuable insights into why an item did not successfully translate since examinees can be directly asked about their interpretation of an item. This advantage can, however, be offset by the interaction between the prober and the examinee being questioned. Cultural, linguistic, and possibly personality differences between the prober and examinee can interfere with the results obtained from the post-translation probe.

A second problem with this method is that it is relatively labor intensive compared to many other judgmental methods. In addition to

enlisting and using probers, examinees are needed to answer test items and respond to probes. Additionally, the probing process is likely to be a time-consuming one.

A third problem with this method is that one has to be sure of the meaning of the answers from source language monolinguals in order to judge the equivalence of the meaning of answers from target language monolinguals. In other words, the validity of the test in the source population must be fully checked before comparing results from source and target examinees. For tests that have not undergone stringent validity checks in the source population (for example, tests that have been developed for small scale research studies), it may be useful to probe a sample of source language monolinguals as well. This sample of monolinguals should be matched as closely as possible to target examinees on the ability or abilities of interest. With this additional check, the problem of comparing irrelevant scores can possibly be avoided.

1.A.2 *Bilingual judges check for errors*. This method makes use of bilingual judges who compare the source and translated versions of each test item and decide whether any differences between translations could result in non-equivalence of meaning in the two populations of interest (Brislin, 1970). These comparisons can be made on the basis of having judges simply look the items over, check the characteristics of the items against a checklist of item characteristics that may introduce non-equivalence, or by having them attempt to answer both versions of the items before comparing them for errors.

One problem in applying this method is that it is often difficult to find bilingual judges who are equally familiar with the source and

target languages and/or cultures. Therefore, judgments about differences between the source and translated versions are subject to variations from this source of error.

A second problem with this method is that bilingual judges may inadvertently use "insightful guesses" to infer equivalence of meaning. This problem is usually raised in the context of using back-translation techniques. Hulin (1987) noted:

> Apparently equivalent terms, such as *amigo*, *friend* and *tovarish*, are not always equivalent, but translators sharing a small number of rules-of-thumb may consistently translate such terms as if they were equivalent. Equivalent source language versions may be generated from poorly translated and constructed target language versions by insightful guesses and assumptions by the translators about what the term must have meant in the original language. Translations that retain grammatical forms of the original language are easy to back-translate but may not be meaningful to target language monolinguals (Brislin, 1970).

Judges are also translators of a sort and are subject to the same errors, in this case using "insightful guesses" to infer equivalence of meaning, as those who performed the initial translation.

A third problem with this method is that bilingual judges may not think about an item in the same way as their respective source and target language monolinguals. Consequently, the use of bilingual judges to establish translation equivalence may lead to results that are not generalizable to source and target language monolinguals. This problem raises serious questions about the overall usefulness of this method for establishing translation equivalence.

1.A.3 *Performance criteria*. This method of establishing translation equivalence is based on the criterion that "if people could perform bodily movements after having heard either a source or target language instructions, and if the results of the bodily movement

criterion were similar across all people, then the source and its translation must be equivalent" (Brislin, 1970). The obvious limitation of this method is that it can only be used with testing materials that can be evaluated through bodily movements such as some test instructions or performance test items. The method has two other problems: It is (1) labor intensive and (2) sensitive to prober-examinee interactions.

1.B.1 *Source language monolinguals check for errors*. Back translation refers to the translation of the target version test back into the source version by bilinguals not involved in the original translation in order to check for translation equivalence (Brislin, 1970). Translation equivalence using this method is established by having source language monolinguals check for errors between the source and back-translated versions of a test (Brislin, 1970; Hulin & Mayer, 1986; Hansen, 1987).

The main problem associated with the use of this method is the reliance on the assumption that errors made during the original translation will not be made again (in reverse) during back-translation. A translator may use "insightful guesses" or "rules-of-thumb" to translate an item, thus making it appear equivalent to the source item even though it may not be. Likewise, the use of these "insightful guesses" and "rules-of-thumb" during the back-translation process can mask those errors made during the original translation. Brislin (1970) reported finding errors due to translation after three successive translation/back-translation sequences, indicating that the assumption that the same errors that occurred in the original translation will not occur, in reverse, during back translation is questionable. The use of additional (independent) translators may make it more likely that

differences in the original translation will be detected, but the high potential for the violation of the previously mentioned assumption reduces the usefulness of this technique and any of the methods discussed that are based on its use.

Therefore, back-translating has problems, but it should be considered a general check on translation quality that will most likely detect obvious errors in the original translation. For example, in an effort to establish translation equivalence of a Spanish translation of the Job Descriptive Index, Hulin, Drasgow, and Komocar (1982) used the back-translation technique as an initial check of translation quality before applying another method of establishing translation equivalence.

Statistical Methods

The three statistical methods to be discussed result from variations in (1) type of examinee responding (source language monolinguals, target language monolinguals, or source-target bilinguals) and (2) version of the test (original, translated, or back-translated) to which the examinees respond. Altogether, four statistical methods will be discussed. To facilitate the discussion of the statistical methods of establishing translation equivalence, the potential statistical techniques used with the three statistical methods will be introduced next.

The statistical techniques used with the various methods of establishing translation equivalence can be categorized along two dimensions: the first dimension is whether it is assumed that the test constitutes a common scale on which scores can be compared. The second dimension is whether the statistical technique conditions on the ability of the examinees being compared. See van de Vijver and Poortinga (1991)

for information on statistical techniques organized by the two dimensions.

Two comments concerning the statistical techniques are in order. First, as van de Vijver and Poortinga (1991) point out, the distinction between the conditional and unconditional statistical techniques is not absolute but rather is dependent on the empirical use of a particular technique:

> . . . the classification of particular techniques as unconditional methods is mainly determined by their empirical use. The [unconditional] methods mentioned can also be applied as conditional methods, namely by including level of ability as an additional factor in the analysis. Suppose a researcher wants to compare p-values obtained in various cultural groups. An unconditional analysis entails a direct comparison of the item statistics, while in a conditional analysis the samples of subjects will be divided according to the level of their raw score and analysed per level. Conversely, the conditional methods which will be discussed, can also be used in an unconditional way by eliminating ability as a separate factor during the analysis (van de Vijver & Poortinga, 1991).

Second, the statistical techniques are often used in combination with a particular statistical method of establishing translation equivalence. For example, to establish the degree of translation equivalence for the English to Spanish translation of the Strong-Campbell Interest Inventory, Hansen and Fouad (1984) used the following statistical indexes in conjunction with method 2.A.1:
(1) Pearson correlation coefficients between group scores on the two forms and (2) dependent samples t-test between mean scores on the two forms.

2.A.1 *Bilinguals take source and target versions*. In this method, bilingual examinees take both the source and target versions of a test (with an adequate time interval in between administrations) and

the scores on the two tests are then compared (Katerburg, Hoy, & Smith, 1977; Hulin, Drasgow, & Komocar, 1982; Hansen & Fouad, 1984; Candell & Hulin, 1986). The source version of the test can either be the original version or a version that has been revised after being checked for translation equivalence with another method. The appeal of this method is that by having the same examinees take both versions of a test, differences in examinee ability that can confound translation equivalence will be controlled for. However, the problem of unequal examinee bilingualism and/or biculturism also applies to the examinees used with this method. The possibility of unequal examinee bilingualism and/or biculturism can violate the assumption of equal examinee ability. Therefore, the assumption that the use of bilinguals controls for differences in ability that would most likely occur if separate source and target language monolinguals were used instead is questionable in many instances.

One way to strengthen this method is to use examinees who are identified as being equally bilingual by a test of language dominance. Several drawbacks with this additional step are evident. These include (1) obtaining or developing a test of language dominance for the source and target languages of interest, (2) the additional required testing time, and (3) the lack of counterpart tests that address biculturism or culture dominance. This additional step may, however, be a practical addition to this method when a test of language dominance appropriate to the source and target languages is readily available.

Another way to strengthen this method is to use statistical techniques that condition on examinee ability. In the few examples provided in the translation literature where this method of establishing

translation equivalence was used, unconditional statistical techniques such as correlations between scores or the use of generalizability theory have been used to compare examinee scores from the source and target versions of the test. These unconditional statistical techniques were used because it was assumed that the use of bilinguals controls for differences in examinee ability. However, as previously mentioned, this assumption is questionable and therefore the use of conditional statistical techniques, such as the use of item response theory (IRT), can be used to strengthen this method of establishing translation equivalence.

Another comment concerning the use of bilinguals in establishing translation equivalence deserves mention. Historically, bilingualism was thought to be a language handicap that interfered with intellectual development and academic achievement (see reviews by Darcy, 1953, 1963). In contrast, recent research in this area (see the review by Diaz, 1983) indicates that compared to monolinguals, bilinguals who are equally proficient in the use of two languages "show definite advantages on measures of metalinguistic abilities, concept formation, field independence, and divergent thinking skills" (Diaz, 1983). Thus, in using bilinguals to establish translation equivalence, the resulting scores may be in general higher than if source and target language monolinguals were used. In the extreme case, floor effects may be noted when the final version of the source and target tests is administered to monolinguals in their respective languages. This problem can arise due to errors in sampling as well, but the use of bilinguals can possibly add a further dimension to this source of error.

The most serious problem with this method, however, is that the scores obtained from bilingual examinees may not be generalizable to their respective source language monolinguals. This problem has been tested empirically by Drasgow and Hulin (1986). They compared previous results of establishing translation equivalence of a Spanish translation of the Job Descriptive Index where bilingual subjects were used (Hulin, Drasgow, & Komocar, 1982) to results using monolingual subjects. In both cases, item response models were used to establish translation equivalence. When bilingual subjects were used (method 2.A.1), approximately 4% (3 out of 72) of the items were determined to have been poorly translated as compared to 30% when monolingual samples (method 2.A.2) were used. Hulin and Mayer (1986) conducted a similar study and obtained similar results. These discrepancies in the number of items identified as poorly translated indicates that the results of establishing translation equivalence based on bilingual responses are likely not generalizable to monolingual populations.

This problem of generalizing results from bilinguals to monolingual populations has been the major reason for the increased interest in method 2.A.2.

2.A.2 *Source language monolinguals take source version and target language monolinguals take target version*. In this method, source and target language monolinguals are used, with each taking the version that is in their respective languages (Candell & Hulin, 1986; Hulin & Mayer, 1986; Hulin, 1987). The source version of the test can either be the original version or a version that has been revised after being checked for translation equivalence with another method. The two sets of scores

are then compared to establish the extent of translation equivalence between the two versions.

The main advantage of this method is that source and target language monolinguals are used and therefore the results of establishing translation equivalence based on this method are more generalizable to these two sub-populations than the statistical methods that use only source language monolinguals (2.B.1) or, to a lesser extent, bilinguals (2.A.1) as examinees. This is due to the concern that bilinguals may not respond to items in the same way that monolinguals in either common source language (see criticisms of 1.A.2, 2.A.1, and 2.B.1). The use of source and target language monolinguals reduces the question of generalizability of the results obtained with this method to the choice of monolingual samples and the statistical techniques used.

The problem with this method is that two samples of examinees are used and therefore the resulting scores may be confounded with differences in ability between the two samples. However, alternative steps can be taken to minimize this problem.

First, in choosing samples of source and target language monolinguals, every effort should be given to matching examinees in the two groups on the ability or abilities or interest. An external criterion such as IQ or other test scores that are correlated with the tasks of interest may be available for this purpose. Alternately, if an external criterion is not available, examinee samples should be chosen using the most available information about the ability level of each sample. Information such as years and type of schooling, age, gender and demographic data may be used for this purpose.

Second, conditional statistical techniques that take into account the ability of examinees when comparing test scores can also be used to control for ability differences in the source and target examinee samples. Examples of conditional statistical techniques that can be used for this purpose include those based on item response models (Hambleton & Swaminathan, 1985; Lord, 1980). The use of item response models are, in particular, receiving much recent attention as a statistical technique used with this method (Hulin, Drasgow, & Komocar, 1982; van der Flier, 1982; Poortinga, 1983; Hulin & Mayer, 1986; Candell & Hulin, 1986; Hulin, 1987; van de Vijver & Poortinga, 1991; Simon, 1989). The advantages of using item response models for this purpose will be discussed in a subsequent section.

Lastly, factor analysis, or other statistical techniques in which no common scale for scores from the populations is assumed, is often used in conjunction with this method to establish translation equivalence (Kline, 1983; Poortinga, 1983; Hulin & Mayer, 1986; van de Vijver & Poortinga, 1991). In the case of factor analysis, scores from source and target language monolinguals are separately analyzed to determine the similarity of factor structures across the populations. The results of a factor analysis are limited in generalizability to similar samples of source or target language monolinguals. This is the case since factor analysis is based on classical item statistics and therefore the results are not sample invariant.

2.B.1 *Source language monolinguals take original and back-translated versions*. In this method, source and back-translated versions are both taken by source language monolinguals and, as with all of the statistical methods, the scores are then compared using one or

more statistical techniques to establish the extent of translation equivalence. The advantage of using this method is that by using one sample of examinees, the resulting scores are not confounded with differences in examinee ability.

One problem with this method is that one set of scores is based on a back-translated version which can mask errors made during the original source to target version translation. An additional problem with the use of this method is that target language monolinguals are not used and yet, in part, the aim is to generalize the meaning of the resulting test scores to a population of target language monolinguals. Making such generalizations without obtaining test scores from at least a sample of the population of interest is a concern with the use of this method.

Review and Selection of Methods

What is evident from the discussion of methods is that certain problems with using the individual methods of establishing translation equivalence are common to several of the methods. In an attempt to provide a basis for choosing one or more methods over others, six problems will be reviewed briefly.

1. Improper to generalize results to the items of interest

   We are ultimately interested in how examinees in the two populations of interest respond to the test items in their respective languages. A problem with method 1.B.1 is that examinees are not required to answer test items (only to check for errors). Since comparing test items for errors in translation may involve different cognitive processes than responding to them, it may be incorrect to generalize from the task of checking for errors in test items to the task of responding to test items.

This problem may also apply to method 1.A.2 when judges are asked only to compare source and target items instead of basing their comparison on their own responses to the items.

2. Improper to generalize results to the populations of interest

A problem with methods 1.B.1 and 2.B.1 is that target language monolinguals are not used and yet it is this population that we are, in part, generalizing the meaning of the resulting test scores to.

The same problem exists for those methods that make use of bilinguals (1.A.2 and 2.A.1). In these methods, the assumption is made that bilinguals will respond to an item in the same way as monolinguals in either language. This is a questionable assumption to make and therefore it may confound the results obtained using these methods. However, the use of bilinguals will most likely be less of a problem in generalizing to the populations of interest than the use of only source language monolinguals.

3. Differences in judges' or examinees' ability

Method 2.A.2 makes use of source and target language monolinguals and therefore the results obtained from this method may be confounded with ability differences between the two groups. This problem also applies to the methods that make use of bilingual judges or examinees (1.A.2 and 2.A.1), although probably to a lesser extent than with the use of source and target language monolinguals. Differences in group or bilinguals' abilities when using methods 2.A.2 or 2.A.1 can be controlled for by the use of conditional statistical techniques. The problem still remains

with method 1.A.2, which uses bilingual judges, since differences in judges' abilities between the source and target languages cannot be controlled for statistically.

4. Use of back-translations

The use of back-translations may cause problems in establishing translation equivalence because errors made in the original source to target translation may be made (in reverse) during the back translation (this may be due to insightful guesses made by the back-translator[s]). Thus, errors made in the original translation may be masked by using those methods that make use of back-translations (1.B.1 and 2.B.1). Back-translating may be useful for picking up obvious errors in the original translation; however, it may not be as useful for picking up more subtle translation errors.

5. Sensitivity to examiner/prober-examinee interaction

All of the statistical methods require administering a test to examinees and, therefore, examiner-examinee interactions may effect the resulting scores. However, the judgmental methods that make use of post-translation probes (1.A.1) or performance criteria (1.A.3) are especially sensitive to examiner/prober-examinee interactions since these methods, in all likelihood, involve a high degree of contact between those administering the test or probes and examinees.

6. Labor intensive

Methods 1.A.1 and 1.A.3 can be relatively labor intensive compared to, for example, having bilingual judges check for errors

(1.A.2). This will be particularly true if a large sample of target language examinees is used.

- -

Insert Figure 2 about here.

- -

These six problems, and the methods of establishing translation equivalence to which they apply, are shown in Figure 2. Besides providing an overview of the problems associated with each method, this Figure can be used to help minimize the errors associated with establishing translation equivalence when more than one method is used. For example, within the judgmental methods, it can be seen from Figure 2 that methods 1.A.2 and 1.B.1 have two problems in common and therefore these two methods should not be used together to establish translation equivalence. A better combination to use may be methods 1.A.1 and 1.A.2 or 1.A.1 and 1.B.1 since these combinations do not share the same problems. Across the judgmental and statistical methods, methods 1.A.3 and 2.A.2 may be a good combination to use for the same reason. Using more than one method will result in a more stringent check of translation equivalence when the methods used minimize the problems they have in common.

However, the choice of method or methods should not be made simply on the number of problems avoided by their use. For one, some problems may be considered more serious than others. For example, budget or time limitations may rule out the use of those methods that are labor intensive (1.A.1 and 1.A.3). Even across methods, the seriousness of a problem may vary. An example is problem 2 (generalizability to the populations of interest), which is most likely a more serious problem when only source language monolinguals (1.B.1 and 2.B.1) rather than

bilinguals (1.A.2 and 2.A.1) are used. External factors can also influence the seriousness of a problem. An example is problem 3, where the seriousness of this problem for the statistical methods (2.A.1 and 2.A.2) varies depending on whether conditional statistical techniques are used with these methods or not. These examples point out that the choice of method or methods used depends on many factors. Figure 2 can provide a frame of reference for considering the various available methods and potentially viable combinations, but the final choice of method or methods used should ultimately be based on further considerations as well.

An additional use for Figure 2 might be to compare judgmental and statistical methods in identifying items that failed to translate well. This has been an important line of research in the study of item bias because identifying why judgmental methods failed to flag the same items as the statistical methods can lead to insights into the nature of item bias. This information can be used by item writers in reducing the number of biased items written and to help in develop better judgmental methods so potentially biased items can be detected before being administered to examinees. Likewise, comparing judgmental and statistical methods in identifying items that failed to translate well can provide comparable information and advantages in the context of translating test items.

Figure 2 can be used when comparing judgmental and statistical methods for flagging poorly translated items by noting the number of problems shared by the judgmental and statistical methods being compared. If the two (or more) methods do not have some problem or problems in common, it would not be surprising to find inconsistent

results across the methods. An example would be comparing the judgmental method 1.B.1 with the statistical 2.B.1. Different problems have been identified across the two methods and therefore consistent results across the methods would appear unlikely from the outset. Similarly, the information in Figure 2 could also be used when comparing across just judgmental or statistical methods. However, the reader is cautioned against interpreting Figure 2 without considering other factors that may influence the seriousness of the problems mentioned.

In summary, seven methods (four judgmental and three statistical) of establishing translation equivalence have been introduced along with a discussion of their respective advantages and problems. With the exception of method 2.B.1, these methods represent the methods of establishing translation equivalence that were found in a review of the relevant literature. Other methods are possible. For example, method 1.A.1 could be extended to include post-translation probes of source language monolinguals who take the source version of a test. Method 1.A.3 could be extended in a similar way, resulting in an additional method of establishing translation equivalence. However, these additional methods are either variations or extensions of the basic methods presented here and, as such, their respective advantages and problems can be evaluated using the discussions presented in this section.

Item Response Models in Establishing Translation Equivalence

Introduction

The discussion earlier highlighted the advantages of using method 2.A.2 (source language monolinguals take source version and target

language monolinguals take target version) for establishing translation equivalence. The main advantage of this method is that translation equivalence results based on its use are more generalizable to the populations of interest (source and target language monolinguals) than with other methods of establishing translation equivalence. The main disadvantage of this method is that these results can be confounded with ability differences between the two samples of examinees. However, these ability differences can be controlled for by applying a conditional statistical technique when comparing examinee responses. Although a number of conditional statistical techniques are available for this purpose, the use of item response models is theoretically preferred when comparing groups of examinees who differ in ability (Hambleton, 1989; Hambleton & Swaminathan, 1985). For this reason and additional reasons, the focus of attention will now shift to the use of item response models in establishing translation equivalence.

The item response models are those that are commonly used in practice for test development, test evaluation, and other testing applications. Two important points about these models are that they are designed for use with (a) unidimensional tests (that is, the test being used measures one dominant trait) and (b) dichotomously scored test data. Item response models that do not require these restrictions have been developed; however, they will not be considered in this paper. For these reasons, the discussion that follows will be based on the commonly used one-, two-, or three-parameter unidimensional logistic models.

Advantages of Using Item Response Models in Establishing Translation Equivalence

The use of item response models has received much recent attention as a statistical technique for establishing translation equivalence (Candell & Hulin, 1986; Ellis, 1989; Hulin, Drasgow, & Komocar, 1982; Hulin & Mayer, 1986; Poortinga, 1983; Simon, 1989; van der Flier, 1982; van de Vijver & Poortinga, 1991). The reason for this attention is that the framework of item response theory provides potential advantages over other conditional statistical techniques when establishing translation equivalence. These advantages can be obtained when an item response model provides a reasonable fit to the test data and include (1) item statistics (parameters) that are independent of the specific sample of examinees used to calibrate the items; (2) examinee ability estimates that are independent of the specific choice of test items used from the calibrated item pool; and (3) examinee ability estimates of known precision. Of particular importance in a translation equivalence study is the first advantage - invariant item parameter estimates.

Invariant item parameter estimates are particularly useful in a translation equivalence study because they provide a strong basis for taking into account differences in examinees abilities when comparing item parameters across populations. Comparisons of item parameters across populations can be carried out by a number of different conditional statistical techniques other than the use of item response models. However, these alternative techniques can be problematic. For example, those methods based on the chi-square statistic are sensitive to sample size and the number of total score intervals used. The Mantel-Haenszel statistic provides a close approximation to results obtained using the one-parameter logistic model but fails to flag items when non-uniform bias is present (Hambleton & Rogers, 1989). When it is

possible to use them, item response models are generally preferred for identifying items that are functioning differently across populations because they (1) explicitly state the relationship between examinee ability and the probability of obtaining a correct response on an item and therefore are a more direct way of identifying differentially functioning items and (2) provide invariant parameter estimates (Mellenbergh, 1983).

It should be noted that invariant examinee ability estimates are also of interest in the context of designing and using translated tests for comparing examinees across populations. When using item response theory in a translation equivalence study, items that did not translate well (non-equivalent items) can be placed on the same ability (or difficulty) scale as those that did translate well (equivalent items). Hulin (1987) noted two benefits of using non-equivalent items when comparing examinees across populations. The first benefit is that instruments can be designed and administered that are potentially more meaningful to the populations of interest:

> The potential for producing equated scales containing mixtures of both emic and etic items offers an additional advantage of IRT procedures in translation and cross-language research. Assuming there are a number of well-translated etic items and that the new emic items meet the assumption of IRT and reflect differences in the same unidimensional latent trait as the culturally general etic items, investigators can tailor scales to each culture by adding a number of emic items specific to each culture to the common set of culturally general etic items. This should increase the sensitivity and cultural relevance of the instrument for both cultures, yet retain the psychometrically required property of equated trait estimates. (Hulin, 1987)

(The term emic refers to terms or concepts that are specific to a population. Its counterpart, etic, refers to terms or concepts that are

universal across populations.) If the items within an instrument are more meaningful to examinees within a population, it is likely that the instrument will also have greater reliability and validity within the population.

The second benefit of using non-equivalent items when comparing examinees across populations is that the precision of examinee ability estimates in each population is increased:

> The presence of many emic concepts in the source language of a particular scale would generate evidence of psychometrically non-equivalent items across the source and target language versions of the instrument. The nonequivalent items could be eliminated and conclusions about θ could be based on the items that were well translated and met the criterion of psychometric equivalence above. However, this involves eliminating the item from both versions of the questionnaire. If the translated item is nonequivalent in the target language but has a nonzero slope for the target language ICC, the item still provides information about θ in both cultures. The information about θ in both languages and cultures provided by the revised scale after eliminating all nonequivalent items would be less than if the entire scale consisting of the complete set of items were scored and used to estimate θ. Cross-cultural comparisons based on more information about θ in both cultures are more precise (Hulin, 1987).

Both of these additional benefits of using non-equivalent items when comparing examinees across populations accrue from invariant examinee ability estimates that can be obtained within the framework of item response theory. Even though these additional benefits are not directly related to establishing translation equivalence (these benefits can only be obtained after completing a translation equivalence study), they offer further compelling reasons for using the framework of item response theory in comparing examinees across populations where differences in language or culture exist.

The advantages of using item response models over other conditional statistical techniques in establishing translation equivalence are gained at a cost. Aside from practical considerations such as the use of large sample sizes and relatively complex numerical procedures, restrictive assumptions about the test, its administration and the resulting scores must be made. These assumptions include (1) test unidimensionality, (2) non-speeded test administration, and (3) an adequate fit of resulting test scores to an item response model (Hambleton & Swaminathan, 1985). Each of these assumptions make it less likely that item response models can be used to establish translation equivalence. However, these assumptions can be checked and, when they are met, the advantages provided by using item response models in cross population comparisons are both unique and extremely useful.

Example of a Translation Equivalence Study

One example of a study to establish translation equivalence will be presented in this last section in order to provide an overview of how the methods of establishing translation equivalence have been used in practice. This example was chosen because it illustrates the use of one of the more popular methods, 2.A.2, of establishing translation equivalence.

The example of a study to establish translation equivalence is Angoff and Cook's (1988) study on the equating of the English and Spanish versions of the Scholastic Aptitude Test (SAT). Their study focused on (1) establishing the translation equivalence for a set of anchor items to be used in equating the two language versions of the SAT and (2) the equating procedure itself. Since we are mainly interested

in the methods and procedures used to establish translation equivalence, the equating portion of this study will not be discussed here.

The first step in establishing translation equivalence was to translate the already existing English version of the SAT into Spanish and the already existing Spanish version into English. The two translated versions were then back-translated into their respective original languages by translators who were not involved in the original translations. Method 1.B.1 was then used to check for errors between the source and back-translated versions for the two language versions of the test. In each case, differences between the source and back-translated versions were noted and either (1) adjustments in the original translations were made if it was determined that the adjustments were adequate to provide potential translation equivalence or (2) the items were dropped as potential anchor items if it was determined that translation equivalence was unlikely to be obtained for these items.

The next phase of this study made use of method 2.A.2. In this case, either the English or Spanish version can be considered the source or target version. After examinee responses from a sample of source and target language monolinguals were obtained, item characteristic curves (ICCs) were estimated separately for each of these groups (the three-parameter logistic model was used). The item parameters were then scaled to allow for comparisons of the ICCs between the two groups. The final set of ICCs for each group were obtained after using a criterion purification procedure developed by Lord (1980, chap. 14). This procedure reduces the problem of using ability and item parameter estimates that may be obtained from non-equivalent items to establish

the equivalence of translated items. The final set of ICCs for source and target language monolinguals was compared to establish the translation equivalence of potential anchor items that were to be used in equating the two language versions of the SAT.

Comparisons of ICCs were based on a combination of indices. First, a chi-squared item bias statistic was calculated for each item. This statistic tests the null hypothesis that the values for the difficulty, discrimination, and pseudo-chance parameters for individual ICCs are the same for the two groups. Items within the verbal and mathematics sections of the test were ranked according to their chi-square values. The mean of the absolute difference between ICCs (Cook, Eignor, & Peterson; 1985) was then calculated for items with relatively small chi-square values. This new difference statistic was used because it, unlike the chi-square statistic, detects differences in ICCs when non-uniform differences are present. From those items with the smallest chi-square values, verbal and mathematics items with smaller mean absolute differences were considered equivalent and used as potential anchor items to equate the two language versions of the test. It should be noted that consideration was given to the language of origin, item type (e.g., antonyms, analogies) for verbal items and content area (e.g., algebra, geometry) for mathematics items when the final set of equating items was chosen.

The Angoff-Cook example illustrates the use of two of the more popular methods of establishing translation equivalence. In the example, a judgmental method (Method 1.B.1) of establishing translation equivalence was used before applying a statistical method for the same purpose. That method 1.B.1 was used is not unusual. Method 1.B.1 is by

far the most common judgmental method of establishing translation equivalence in use today and is used almost routinely as a general check of translation equivalence.

The example also illustrates the use of one of the currently popular statistical methods of establishing translation equivalence (2.A.2). The use of method 2.A.2 is, however, a more recent trend due to the established feasibility of using item response models in conjunction with this method.

References

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (Report No. 88-2). New York, NY: College Entrance Examination Board.

Berry, G. L., & Lopez, C. A. (1977). Testing programs and the Spanish-speaking child: Assessment guidelines for school counselors. The School Counselor, March, 261-269.

Brislin, R. (1970). Back-translation for cross-cultural research. Journal of Cross-Cultural Psychology, 1, 185-216.

Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology, 17(4), 417-440.

Cook, L. L., Eignor, D. R., & Petersen, N. S. (1985). A study of the temporal stability of item parameter estimates (ETS Research Report 85-45). Princeton, NJ: Educational Testing Service.

Darcy, N. T. (1953). A review of the literature on the effects of bilingualism upon the measure of intelligence. Journal of Genetic Psychology, 82, 21-57.

Darcy, N. T. (1963). Bilingualism and the measure of intelligence: Review of a decade of research. Journal of Genetic Psychology, 103, 259-282.

DeAvila, E. A., & Havassy, B. (1974). The testing of minority children - a neo-Piagetian approach. Today's Education, December, 72-75.

Diaz, R. M. (1983). Thought and two languages: The impact of bilingualism on cognitive development. In E. W. Gordon (Ed.), Review of research in education, Volume 10. Washington, DC: American Educational Research Association.

Drasgow, F., & Hulin, C. I. (1986). Assessing the equivalence of measurement of attitudes and aptitudes across heterogeneous subpopulations. Unpublished manuscript. University of Illinois at Urbana-Champaign.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74, 912-921.

Fouad, A. N., & Hansen, J. C. (1987). Cross-cultural predictive accuracy of the Strong Campbell Interest Inventory. Measurement and Evaluation in Counseling and Development, 20, 3-10.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed.; pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of the IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Hansen, J. J. (1987). Cross-cultural research on vocational interests. Measurement and Evaluation in Counseling and Development, 19, 163-176.

Hansen, J. C., & Fouad, A. N. (1984). Translation and validation of the Spanish form of the Strong-Campbell Interest Inventory. Measurement and Evaluation in Counseling and Development, 16, 192-197.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18, 115-142.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. Journal of Applied Psychology, 67, 818-825.

Hulin, C. L., & Mayer, L. M. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. Journal of Applied Psychology, 71(1), 83-94.

Katerburg, R., Hoy, S., & Smith, F. J. (1977). Language, time and person effects on attitude scale translations. Journal of Applied Psychology, 62, 385-391.

Kline, P. (1983). The cross-cultural use of personality tests. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Lapointe, A. E., Mead, N. A., & Phillips, G. W. (1989). A world of differences: An international assessment of mathematics and science (Report No. 19-CAEP-01). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Mellenbergh, G. J. (1983). Conditional item bias methods. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Olmedo, E. L. (1981). Testing linguistic minorities. American Psychologist, 36, 1078-1085.

Poortinga, Y. H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Samuda, R. J. (1983). Cross-cultural testing within a multicultural society. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Simon, M. G. (1989, March). Bias in translated items: A comparative study of five statistical methods. Paper presented at the meeting of AERA, San Francisco.

Stanley, J. C., & Hopkins, K. D. (1972). Educational and psychological measurement and evaluation. Englewood Cliffs, NJ: Prentice Hall.

Swanson, H. L., & Watson, B. L. (1982). Educational and psychological assessment of exceptional children. London: C. V. Mosby Company.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. Journal of Cross-Cultural Psychology, 13, 267-298.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Culture-free measurement in the history of cross-cultural psychology. In R. K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing. Boston: Kluwer Academic Publishers.

Figure 1. Methods for Establishing Equivalence of Translated Test Items

1. Judgmental Methods

1.A Judgmental single-translation methods

	Source	Target
1.A.1.	---------->	Post-translation probes
1.A.2	---------->	Bilingual judges check errors
1.A.3	---------->	Performance criteria - perform a task using translated instructions

1.B Judgmental back-translation method

1.B.1 ---------->
<----------

Source language
monolinguals
check for errors

2. Statistical Methods

2.A Statistical single-translation methods

	Source	Target
2.A.1	---------->	Bilinguals take source and target versions
2.A.2	---------->	
	Source language monolinguals take source version	Target language monolinguals take target version

2.B Statistical back-translation method

2.B.1 ---------->
<----------

Source language
monolinguals
take source and
back-translated
versions

Figure 2. Problems Associated with the Methods of Establishing Translation Equivalence

Problem	Methods of Establishing Translation Equivalence						
	Judgmental Methods				Statistical Methods		
	1.A.1	1.A.2	1.A.3	1.B.1	2.A.1	2.A.2	2.B.1
1. Improper to generalize results to the items of interest		X		X			
2. Improper to generalize results to the populations of interest.		X[1]		X	X[1]		X
3. Differences in judges' or examinees' ability		X			X[2]	X[2]	
4. Use of back-translations				X			X
5. Sensitivity to examiner/ prober-examinee interactions	X		X		X[3]	X[3]	X[3]
6. Labor intensive	X		X				

An X indicates the problem is associated with the method.

[1]Most likely less of a problem than using only source language monolinguals.

[2]Less of a problem if conditional statistical techniques are used.

[3]Most likely less of a problem than with using probes or performance criteria.