ED 358 128 TM 019 910

AUTHOR	Hambleton, Ronald K.
TITLE	Translating Achievement Tests for Use in Cross-National Studies.
SPONS AGENCY	International Association for the Evaluation of Educational Achievement, New York, NY.; National Center for Education Statistics (ED), Washington, DC.
PUB DATE	[Mar 93]
NOTE	32p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Cross Cultural Studies; *Cultural Differences; Data Collection; Foreign Countries; Guidelines; *International Studies; Interpreters; Language Skills; Quality Control; Research Design; Scores; Second Languages; *Test Format; Testing Problems; Test Use; Test Validity; *Translation
IDENTIFIERS	Test Equivalence

ABSTRACT

Translating achievement tests and questionnaires prepared in one language and culture for use in other languages and cultures has been a long-standing practice. Unfortunately, there is considerable technical evidence that suggests that the quality of test translations varies considerably, and too often the translations are not very good, thus reducing the validity of any results produced with the translated tests and questionnaires. Four important problems that arise in translation that must be considered are: (1) the selection of translators; (2) identifying the appropriate language for the target version of the test; (3) identifying and minimizing cultural differences; and (4) finding equivalent words or phrases. Several methods for establishing the equivalence of translated tests are reviewed. Two judgmental methods are forward translation (pretest method) and back translation, in which the test is translated to the target language and back to the original by a second translator. Three data collection designs that have been used in establishing test score equivalence of source and target language versions of a test are reviewed. Fourteen guidelines are presented for translating tests and establishing score equivalence. A 40-item list of references is included, as well as French and English abstracts of the document. (SLD)

Translating Achievement Tests for Use in Cross-National Studies

Ronald K. Hambleton
University of Massachusetts at Amherst, U.S.A.

Abstract

Translating achievement tests and questionnaires prepared in one language and culture for use in other languages and cultures has been a longstanding practice. Unfortunately, there is considerable technical evidence which suggests that the quality of test translations vary considerably and too often the translations are not very good, thus reducing the validity of any results produced with the translated tests and questionnaires. The specific purposes of this paper are (1) to address four important problems that arise when translating tests and how these problems might be resolved, (2) to identify and review judgmental and statistical methods for establishing the equivalence of scores from the same test presented in different languages, and (3) to provide some preliminary guidelines for persons doing test translations and equivalence studies.

Translating Achievement Tests for Use in Cross-National Studies[1,2,3]

Ronald K. Hambleton
University of Massachusetts at Amherst, USA

Translating achievement tests and questionnaires prepared in one language and culture for use in other languages and cultures has had a long history in educational and psychological testing. One of the earliest examples was the translation of the Binet-Simon *Intelligence Scale for Children* from the French language into the English language in 1911. By 1916 the Binet-Simon intelligence test had been translated into seven languages. Today, the practice of translating tests and questionnaires from one language and culture to others is widespread (Oakland & Hu, 1992). Tests of intellectual ability such as the *Wechsler Intelligence Scale for Children* and the *Stanford-Binet Intelligence Scale* and measures of personality such as the *Thematic Apperception Test* and the *Minnesota Multiphasic Personality Inventory* are among the instruments that have been most frequently translated (Oakland & Hu, 1992).

There appear to be at least two reasons for translating tests. First, the practice is economical: For those who want to assess a particular construct, it may be far less expensive and considerably faster to translate a test or questionnaire than to construct a new instrument to measure the construct of interest in a second language. Sometimes, too, the technical

[1] Support for the preparation of this paper was provided by the International Association for the Evaluation of Educational Achievement (IEA) and the National Center for Educational Statistics (NCES) of the United States Department of Education. The opinions here, however, are those of the author and not necessarily those of IEA or NCES.

[2] *Laboratory of Psychometric and Evaluative Research Report No. 241*. Amherst, MA: University of Massachusetts, School of Education.

[3] To appear in the *European Journal of Psychological Assessment*.

expertise is simply not available in the second language to construct the needed instrument. This is the reason popular American and British psychological tests have been translated into many languages.

A second reason for translating tests and questionnaires from one language to another is to permit cross-national studies (see, for example, Miura, et al., 1993). Such studies have been conducted in the areas of educational achievement and school attitudes for over thirty years by The International Association for the Evaluation of Educational Achievement (IEA) (see, for example, Keeves, 1992). These studies have taken on special importance for educational policy-makers in recent years. In the United States, for example, policy-makers want to set 'world-class educational standards.' Achievement test results which permit valid comparisons between American students and students from other countries would be valuable information to American educational policy makers in the standard-setting process. Other countries have similar reasons for being interested in cross-national studies.

As evidence of the increasing importance of cross-national studies of achievement, one needs only know that in excess of 60 countries are planning to participate in IEA's Third International Mathematics and Science Study (TIMSS) scheduled for 1994 or 1995. Previous IEA studies have had substantially smaller numbers of participating countries. Also too, the number of cross-national studies has been on the increase. For example, The United States Department of Education sponsored studies of mathematics and science achievement in 1988 and 1991 involving five and twenty countries, respectively (Lapointe, Mead, & Phillips, 1989; Lapointe, Mead, & Askew, 1992).

Continuing interest in translating tests and questionnaires can be expected in the coming years. China, India, and many countries in Africa are already heavy users of translated tests from the United States and Great Britain. As for personality measures, Spielberger's instrument to measure trait-state anxiety (see Spielberger, et al., 1971) has been translated into over 40 languages and the number has been steadily increasing (personal communication with Charles Spielberger, September 27, 1992). Political, social, and economic changes in Europe will impact, too, on educational testing and the need to translate tests.

With respect to cross-national studies of achievement, there are many threats to the validity of the interpretations of the findings including (1) selection of comparable samples, (2) test administration conditions, and (3) equal familiarity of the testing formats (e.g., multiple-choice test items). At least as important as these, is the threat of an improper translation. Unless the translation work is done well, and evidence is compiled to establish the psychometric equivalence of the two or more versions of a test, questions about the validity of any uses of the translated tests should arise. Also, the validity of achievement comparisons among countries where different versions of the test were administered will be in doubt until questions about the equivalence of the versions of the test are resolved.

The general purpose of this paper is to review issues and methods associated with translating achievement tests. The specific purposes are (1) to address four important problems that arise in translating tests and how they might be solved, (2) to identify and review several judgmental and statistical methods for establishing the equivalence of scores from the same test presented in different languages and cultures, and (3) to provide several preliminary guidelines to those who are doing test translations and equivalence studies.

Some researchers prefer the term test adaptation to test translation because the former term seems to more accurately reflect the process that often takes place: Producing an equivalent test in a second language or culture often involves not only a translation that preserves the original test meaning but also additional changes such as those affecting item format and testing procedures may be necessary to insure the equivalence of the versions of the test in multiple languages or cultures. In this paper, the term test translation will be used because it is familiar to many researchers. But it will be used in the broader sense to include all changes that may be necessary to produce the desired results, i.e., equivalent versions of a test in two languages and cultures.

Attention in the paper is focused on translating achievement tests though most of the discussions that follow apply equally well to tests and questionnaires in the areas of personality and attitude measurement. Unique problems and methods associated with translating personality inventories, attitude scales, and questionnaires will not be considered here. Readers are referred to papers by Spielberger, et al. (1971) and Jackson (1991) for more information on these topics.

ADAPTING TESTS: FOUR COMMON PROBLEMS AND POSSIBLE SOLUTIONS

Four problems along with possible solutions will be considered in this section: the selection of translators, identifying the appropriate language for the target version of the test, identifying and minimizing cultural differences, and finding equivalent words or phrases.

The Selection of Translators

Perhaps surprisingly, one of the common problems involves the selection of translators. The task of choosing translators seems straightforward enough: Find persons with an excellent knowledge of the source and target languages. However the technical literature suggests that at least three other qualifications are necessary. First, successful translations are carried out by persons who are knowledgeable about the subject matter. In one recent translation of the technical term "item pool" the translation into Japanese was "item oceans." This example highlights the shortcomings with a literal translation. Technical knowledge on the part of translators is essential or the meaning of the source material can easily be lost in the translation (Brislin, 1970). Second, successful translations are carried out by translators who have experiences in both languages. Experienced test translators such as Woodcock (1985) recommend however when translating from the source language to the target language that those involved with the project should be dominant in the target language and have experiences in that culture. Otherwise, it is often very difficult to achieve a satisfactory translation. According to Woodcock (personal communication, May 9, 1992), "Few persons, for whom the target language has been acquired later, will be as sensitive to the unique patterns of a language that, when present, makes a translation sound natural and not stilted."

Finally, test translations are done best by persons who have skills in test development, and know the principles of writing good test items. These skills are essential so that common errors in item writing do not ente. during the translation process. For example, a translator not familiar with multiple-choice test item writing could introduce "clang" associations, unusually long correct answers, distractors that have the same meaning, awkward item stems, etc., that reduce the validity of the test items in the target language. Some of the errors could make the test items harder (e.g., awkward item stems) and other errors might make the test items easier (e.g., two or more distractors with the same meaning or value). The result, however, is the same: a non-equivalent test.

Identifying the Appropriate Language for the Target Version of the Test

A problem in translating tests sometimes arises because of multiple dialects within the target language (Olmedo, 1981). According to Olmedo:

> ...it is not uncommon to find that many tests written in formal Spanish are used inappropriately with populations that speak substantially different Spanish dialects.

One solution to the problem is to insure that the test is translated into as many dialects or cultural groups as necessary so that examinees are not placed at a disadvantage. In the extreme, each dialect group is treated as if it were a different language group. This solution is not very practical and not usually necessary as will be seen below. Even if this solution is adopted, a problem remains. Examinees must be correctly assigned to the dialect version of the test that would be most familiar to them. But, DeAvila and Havassy (1974) point out that just because a person speaks a language, it cannot be assumed that he or she can read and therefore should be tested in that language. Problems in this identification would need to be dealt with and special problems that may arise due to examinees in the same test setting being administered different versions of the test would need to be handled.

Alternatively, and certainly more efficiently, perhaps a single translation acceptable to all of the dialects within a language can be produced. Woodcock (1985), for example, has done this successfully in Spanish with the Woodcock-Johnson Psycho-Educational Battery. The following excerpt from his technical manual is informative:

> During test development special attention was directed toward designing item and test instructions that would be deemed

> appropriate across the Spanish-speaking world. Thus, professionals from several regions of the Spanish-speaking world were involved cooperatively in item development and the preparation of test instructions. In addition, the publisher established a board of four consulting editors to review and advise on all aspects of the project including the item content and Spanish language text... During the norming of the [instrument] approximately 30 examiners from five Spanish-speaking countries were trained to gather data. Each of the examiners was also responsible for critically reviewing the test text and answer keys for possible Spanish-language problems based on their regional perspective. (Woodcock, 1985, p. 2)

Woodcock's approach to translating his tests into Spanish and dealing with the problem of dialects was based on three principles:

1. The original translation is done by several Spanish-speaking professionals. The test is checked independently by several reviewers and then the reviewers meet with the translators to discuss problems in the translation and attempt to achieve a consensus about the necessary revisions.
2. A translation review team made up of representatives from the different regions of the Spanish-speaking world meets to check the Spanish version of the test prepared at step 1.
3. At the field-test stage, test administrators compile lists of translation and scoring problems that arise.

Results from step 3 combined with some statistical studies such as those described in the next section of the paper are used to prepare the final version of the test. It remains to be demonstrated, however, how well the three steps above which appeared to work well in producing a Spanish translation will work to address dialect problems in other language and cultural groups.

Identifying and Minimizing Cultural Differences

There are a number of cultural factors that will cause problems if they operate differently in the source and target languages. Van de Vijver and

Poortinga (1991) pointed out several 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.

Among the factors that may cause problems are levels of test motivation, unfamiliar test item formats, variable experiences and values, test anxiety, and test speededness (van de Vijver & Poortinga, 1991).

Cross-cultural researchers have provided numerous examples of how cultural variables can impact on test performance. The use of multiple-choice items, for example, may be a problem. At least some cultural and language groups will be substantially less familiar than others with multiple choice test items. Outside the United States, the multiple-choice item format is <u>not</u> common. Perhaps the use of practice materials on multiple-choice items can provide at least a partial solution to the problem of differential familiarity. Another response would be to insure that <u>multiple</u> item formats are used in the test and that analyses are conducted after the test administration to determine the extent of the problem associated with the use of unfamiliar item formats. When comparing the level of achievement of two countries, the problem of differential item familiarity might be suspected if the size of the achievement difference is greater on (say) the multiple-choice items than the essay items. Studies to investigate the seriousness of other test factors such as test speededness might also be carried out.

<u>Finding Equivalent Words or Phrases</u>

A fourth problem that arises in test translations is finding words or phrases that are equivalent in the source and target languages. For reasons of test score validity, every effort must be made to preserve the original meaning of the test directions and items. Sometimes the item substitution method is used. Here, an item which may not translate well (e.g., a source language test item about the President as head of a government may be less meaningful in a country with a prime minister) is replaced by a comparable item. Even this seemingly straightforward procedure is not without problems. In one recent translation/adaptation project in Canada, the effects of "Canadianizing" an American achievement test resulted in a more difficult Canadian version than the original American test had been!

In international comparative studies, a second and powerful alternative exists to the problem of finding equivalent words or phrases. 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 the source version of a test at the test development stage or later, in both language versions of the test, in order to achieve item equivalence. One example comes from a paper by Swanson and Watson (1982). The Spanish word "paloma" is equivalent to either "dove" or "pigeon" in English and therefore a test item in English which required the student to make a distinction between a dove and a pigeon would be difficult if not impossible to translate into Spanish. The original item in English could be decentered by using a pair of terms that do have similar meanings within the context of the item, and do have equivalent terms in Spanish. Thus the change in the original item would permit a correct translation to be made. In international comparative studies of achievement, it would appear that decentering should be a common practice. Brislin (1970) reported that the best translations result when decentering is used.

Two additional points seem worthy of mention. First, difficult to translate words and phrases should be kept in mind and avoided at the test development stage (Brislin, 1986). Second, decentering can be most effectively done after a back translation of a test is prepared. At this stage via a comparison of the original and back-translated test, difficult to translate words and phrases can usually be identified.

Decentering is not without some potential risks to test validity in the two or more languages where it is used. Hulin and Mayer (1986) point out:

> 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.

It is difficult to get a sense of the extent and appropriateness of decentering used in specific test translations from the literature. Typically all that is reported is that decentering was used or it wasn't. Validity evidence to support the use of translated tests would be enhanced if test developers reported the percent of time decentering was done along with illustrative examples.

When the intent from the beginning of a testing project is to produce tests that can be used in multiple languages and cultures such as in TIMSS, at least one other option exists. Recognizing that there may always be problems in translated tests, some test development projects attempt to distribute the problems so that the target language group is not always placed at a disadvantage. For example, in Canada, there are equivalent French and English versions of several credentialing examinations. Half of each exam is prepared in French and translated into English. The other half of each exam is prepared in English and translated into French. Such a test development strategy seems fair because what problems result from test translations are then equally present in both versions.

Summary

Four common problems associated with translating tests have been reviewed and in each case, suggestions were offered for how the problems might be addressed in practice. The extent to which the four problems occur in practice depends upon many factors including the test format, test content, test difficulty, the particular language and/or cultural groups involved, the expertise of the test developers and translators, and the amount of verbal load in the test. Brislin (1970) reported, for example, (1) the languages involved can greatly influence the difficulty of the test translation process-the more similar the structure (e.g., English and French are more similar than English and Chinese) the better the translation, (2) the technical knowledge of the translators is an extremely important factor, and (3) translations tend to be better if translators are given practice and feedback before they begin the task. When the cultures of the groups differ substantially, translating a test for equally valid uses in each becomes an even more complex process. However, being aware of the problems and possible solutions described in this section should improve the quality of test translations.

METHODS FOR ESTABLISHING THE EQUIVALENCE OF TRANSLATED TESTS

Equivalence of test items in the source and target languages means that scores derived from the groups taking each are comparable. Any item and test score differences are due to real differences in proficiency and not to one group or the other being at a disadvantage because of the choice of vocabulary, stimulus material, item format, test directions, etc. It is possible to define equivalence of test items across languages/cultures within the framework of item bias: two versions of an item when prepared in different languages are assumed to be equivalent when members of each group of the same ability have the same probability of success on the item (Hambleton, Swaminathan, & Rogers, 1991). If the probabilities are different, the item is labelled "potentially biased." Other definitions have also been proposed in the literature (see, for example, Brislin, 1970).

There is also a similarity in the psychometric methods used to establish translation equivalence and to identify potentially biased items. In each case, both (1) judgmental methods and (2) statistical methods, may be used. Unfortunately, at least up to 1970, there was little evidence that researchers paid much attention to any of the judgmental and statistical methods in the literature (Brislin, 1970). In his review of 80 research articles in cross-cultural research, Brislin felt 61 studies reported so little information about the test translation process that problems in test translation could not be ruled out. Disappointingly, many of the remaining researchers noted that "a bilingual friend did the translation." Brislin felt that these findings were so significant that they cast doubt on the validity of significant portions of cross-cultural research studies up through 1970. To the extent that the test translation process has not improved substantially in the last 20 years, similar validity concerns might be raised about the more recent cross-cultural literature. Even in the otherwise technically sound international comparative studies of Lapointe, Mead, and Phillips (1989) and Lapointe, Mead, and Askew (1992), only modest attention was given to the technical problems of test translation and establishing test score equivalence. Each participating country was made responsible for doing its own translations. No standardized and validated procedures were prepared to guide the translation process.

In the remainder of this section, a review of several major judgmental and statistical methods for investigating the equivalence of source and target versions of a test will be presented.

Judgmental Methods

Two basic judgmental methods (with variations) were identified in the educational and psychological literature: Forward-Translation and Back-Translation.

Forward-Translation

One variation of this method (sometimes called the "pretest method") is that either a single translator or a group of translators prepares a translation of the source language version of the test into the target language. One or more samples of target language examinees answer the translated version of each item and are asked about the meaning of each item and their answers. Evidence of translation equivalence is obtained when the responses given by a high percentage of the examinees reflects a reasonable interpretation of the item. The main judgment to be made is whether target language examinees perceive the meaning of each test item in the same way as the source language examinees.

The use of this variation of the forward-translation method can provide valuable insights into why an item does not successfully translate since examinees can be directly asked about their interpretations. This advantage, however, is offset by several problems. First, there is the possibility of a failure to communicate between the test translator and the examinees, especially if the test translator is of a different culture and predominant language group than the examinees. A second potential problem is that this method can be labor intensive. A third problem is that the test translator has to be sure of the meaning of the answers from source language examinees in order to judge the equivalence of the meaning of answers from target language examinees. Therefore, it is necessary to conduct intensive investigations of the source language version of the test first. To conduct this type of comparative study correctly, the two samples of examinees should be matched as closely as possible on ability and preferably be representative of the groups being assessed in each language/culture.

The temptation might be to use a single group of bilingual examinees and have them comment on both versions of the test with the use of a design that controls for the order of test presentation effect. At least two problems are present with this design: First, bilingual examinees may, on the average, be different from unilingual examinees in some important way that affects test performance. For example, perhaps the bilingual sample may be generally more capable than the unilingual sample. It is possible, then, that findings cannot be safely generalized to unilingual examinees in each language group. Second, there is the possibility that the bilingual examinees are not equally proficient in both languages. If differences exist, the findings of the comparative study will be difficult to interpret. Language dominance tests could be used but their validity would need to be substantiated as well.

Back-Translation

This method is the best known and most popular of the judgmental methods. A test developer prepares the original version of the test in the source language. In one variation, two bilingual translators are hired to work independently: The first one translates the test from the source language to the target language. Then the second one translates the test from the target language back to the source language. Finally, the two versions of the test in the source language are compared to evaluate the quality of the translations. Of course variations on the basic method are possible: multiple translators can be hired so that teams can do the two translations necessary in the method (referred to in the literature as the "committee method"). Also, translators might be identi.ᶜied with content expertise and instructed to translate the test as seems appropriate to insure equivalent forms in the two languages.

Numerous criticisms have been leveled at this method. For one, the evaluation of test equivalence is carried out in the source language. It is quite possible that the findings in the source language do not generalize to the target language version of the test. Possibly translators use a shared set of translation rules that insures that the back-translated test looks like the original test but little is known about the comparison of interest, i.e. the source and target versions of the test. Brislin (1970) offered an example: He noted that the words "amigo" and "friend" are not always equivalent in Spanish and English. But, if translators share the common convention that they are similar, problems in the Spanish translation are masked.

Another possibility is that the back translator(s) is able to do a good translation even though the original translation was poorly done and resulted in a non-equivalent target language version of the test. For example, the original translation may be poor because it retains inappropriate aspects of the source language test such as some grammar and spelling. Such errors facilitate back-translations but they mask serious shortcomings in the target language version of the test. Finally, as is true with all judgmental methods, no examinees ever see the two versions of the test under true testing conditions and therefore since examinees are often operating at different cognitive levels than the translators, it is not improbable to think that translations found to be acceptable by test developers may not actually be so in practice. A good example of this point is found in the item bias literature. A common finding in the item bias literature is that judgmental methods (which use the opinions of experts) and statistical methods (which are based on the actual item responses of examinees) rarely converge on the same set of flawed items (see, for example, Hambleton & Jones, in press).

One variation involves the use of bilingual translators or judges who check for errors in meaning in the two or more versions of the test. 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. These comparisons could 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 attempt to answer both versions of the items before comparing them for errors. Many of the same problems which apply to forward-translations surface again: (1) it is often difficult to find judges who are equally proficient in both languages and cultures, (2) judges often use insightful guesses in translating from one version of the test and back again but examinees do not share the same experiences, and (3) bilingual translators do not necessarily think about test items in the same way that unilingual examinees might.

Clearly, the back-translation method (and variations) has problems but the method could be considered a general check on translation quality that will detect at least some of the problems associated with poor translations or adaptations. Hulin, Drasgow, and Komocar (1982) used the back-translation method successfuly as an initial check of translation quality before applying a statistical method of establishing test equivalence.

Statistical Methods

Three data collection designs have been used in establishing test score equivalence of the source and target language versions of a test. The designs result from variations in two factors: (1) type of examinee responding (source language monolinguals, target language monolinguals, or bilinguals), and (2) versions of the test administered (original version, translated version, or back-translated version). A description of each design follows along with several evaluative comments.

Bilingual Examinees Take Source and Target Versions

On the surface this design seems reasonable: Bilingual examinees are located and administered the source and target versions of the test. Care is taken, or should be at least, to insure that the order of test presentation is counter-balanced with half the bilinguals taking the versions in one order and the other half taking the versions in the reverse order. Time between administrations should be minimal to insure that ability changes do not take place between administrations. The appeal of this method is that by having the same examinees take both versions of the test, differences in examinee ability that can confound test translation equivalence studies can be controlled. Of course the main flaw is that the premise is potentially faulty. Unless evidence is compiled to show equal proficiency of examinees in each language and/or culture, the design cannot be safely used. One useful variation which suffers from the same flaw but which is more administratively convenient is to split the available sample of bilingual examinees randomly into two groups. Each group is assigned to take only *one* of the tests. Equal ability groups are assumed so that item statistics and the correlational structure of the items can be compared to detect potentially poorly translated/adapted test items. The plausibility of the equivalent groups assumption is one of the keys to the viability of this design.

Analyses that do not require the assumption of equal abilities in the two languages can still be used with this design. For example, the similarity of the rank orderings of item difficulties can be checked. Checks on the factorial invariance across the two groups is another possibility (Joreskog, 1971; Joreskog & Sorbom, 1986). But even with these analyses, threats to the validity of data interpretations are present. As bilingual examinees may tend to be on the average more capable than their monolingual counterparts, the finding of test equivalence may not generalize to the intended populations of examinees in each language group. Historically, bilingualism was thought to be a language handicap that interfered with intellectual development and academic achievement (see, for example, Darcy, 1963). More recently, however, researchers such as Diaz (1983) have found that compared to monolinguals, bilinguals who are equally proficient in the use of two languages show definite advantages on measures of meta-linguistic abilities, divergent thinking, and several other cognitive skills. Thus in using bilinguals to establish test translation equivalence, the resulting scores may be in general higher than if source and target language monolinguals are used. The result is that findings may not generalize to monolingual examinees.

About the only way to salvage anything useful from this type of design is to only use examinees who are identified as equally proficient in both languages by a language dominant test. This approach is sound in theory but has many shortcomings in practice. First, there is a shortage of valid language dominance tests and they exist in only a few languages. Second, the use of additional tests will require more testing time and possibly reduce the number of examinees willing to participate in the study. Finally, there is a shortage of tests that address biculturalism or culture dominance.

In sum, there is little evidence available to support this design. Undoubtedly the most serious problem is that the scores obtained from bilingual examinees may not be generalizable to source and target language monolingual examinees. This problem was investigated empirically by Drasgow and Hulin (1986). They compared previous results of establishing score equivalence of a Spanish translation of the Job Descriptive Index where bilingual examinees were used (Hulin, Drasgow, & Komocar, 1982) to the results obtained using monolingual examinees. (Item response models were used in both studies to identify problem items.) When bilingual examinees were used, about 4% of the items were identified as being poorly translated. The result jumped to 30% of the items when monolingual samples in the target and source languages were used. The discrepancy in results provides rather powerful evidence that the results of establishing translation equivalence based on bilingual responses are not always generalizable to monolingual populations.

Source Language Monolingual Examinees Take the Original and Back-Translated Versions of the Test

This design involves the administration of the original and back-translated versions of the test to a sample of monolingual examinees in the source language. Counter-balancing the order of test administrations is essential. One variation is to randomly assign a group of source language persons to take either the original or back-translated version of the test. This design (and variation) has some merit: For one, since the same group of examinees (or randomly equivalent groups) takes both forms of the test in the same language, the assumption of equal ability is plausible. Also, when examinees show substantially different performance on two versions of an item, possible problems in the translation are identified. But the shortcomings of this design far outweigh the modest advantages. The main shortcomings are that no empirical data is collected on the translated version of the test and predictions of problems that might arise with the translated version of the test must surely be incomplete. Little more probably needs to be said about this very weak design.

Source Language Monolinguals Take Source Version and Target Language Take Target Version

In this method, source and target language monolinguals are used, with each group taking the version that is in their own language. Excellent applications of this design in practical work are provided by Ellis (1989, 1991), Ellis and Kimmel (1992), Candell and Hulin (1986), Hulin (1987), and Hulin and Mayer (1986). The source version of the test could either be the original version or the back-translated version, if the latter version exists. The sets of scores from the two monolingual groups are then compared to determine the equivalence of the versions.

The main advantage of this method is that source and target language monolinguals are used and therefore any findings are more generalizable to the two populations of interest than the results obtained from the other two designs described previously. The use of source and target language monolinguals reduces the question of generalizability of the results obtained to a consideration of the choice of monolingual samples and the statistical methods used in the analyses of the data.

The main problem with this method is that two different samples of examinees are used and it is not appropriate to assume that they are equivalent in ability. In fact, very often the original test is being translated to permit a meaningful comparison of ability differences of samples of examinees in the two language groups. We note that any applications of statistical methods that require an assumption of equal ability differences will produce misleading results. For example, it is common with this design to compare item p values in the two groups but such an analysis is apt to overidentify problematic items. Fortunately, several reasonable steps can be taken.

First, in choosing samples of source and target language monolinguals an effort can be made to match examinees in the two groups on the ability or abilities measured by the test. An external criterion such as an IQ test or another test that is correlated with the test of interest might be used. Also the groups might be matched on sex, age, grade level, and other pertinent demographic variables. But in fact attempting to match samples closely is probably not the preferred direction for addressing the problem. A practical solution is to expect ability differences, though steps can be taken to reduce the size of the differences in the design, and use conditional statistical techniques for comparing item and total score performance in the two groups that take into account any ability differences between the groups.

Examples of conditional statistical techniques that can take into account group differences when making important item comparisons across monolingual groups include item response models (see Hambleton, 1989, or Hambleton, Swaminathan, & Rogers, 1991, for introductions to item response theory and related models) and the Mantel-Haenszel procedure for item bias detection studies (Holland & Thayer, 1988; Hambleton & Rogers, 1989). These new techniques are receiving considerable attention from researchers working in the field of test translations (see, for example, Angoff & Cook, 1988; Ellis, 1989, 1991; Hulir. 1987; Hulin, Drasgow, & Komocar, 1982; van de Vijver

& Poortinga, 1991). These and other promising methods and procedures for implementing the methods may be found by reviewing the item bias literature (Scheuneman & Bleistein, 1989).

Finally, factor analysis, or more generally, covariance structural analysis (for example, as represented by the work of Joreskog, McDonald, Muthen, and others), can be used in conjunction with other methods. In the case of factor analysis, scores from source and target language monolinguals are separately analyzed and then the factor structures obtained in the two samples are compared (Joreskog, 1971; Muthen & Christoffersson, 1981). Similar structures in the two groups provide evidence of the equivalence of the two versions of the test. Non-equivalent structures may suggest problems in the test translation process.

GUIDELINES FOR TRANSLATING TESTS AND ESTABLISHING TEST SCORE EQUIVALENCE

General guidelines for conducting international comparative studies in educational achievement are now available (Bradburn & Gilford, 1990). In view of the fact that expectations are often high for results from cross-cultural or international comparative studies, or for the utility of translated tests, the need for professionally developed and validated technical standards for translating tests and establishing test score equivalence seems clear as well. Disappointingly, the technical literature on these points is rather incomplete (from a measurement perspective), and what literature there is, is scattered throughout a plethora of journals, reports, and books (see, for example, Brislin, 1986; Batcher & Garcia, 1978; Gross & Scott, 1989; Hambleton & Bollwark, 1991; Poortinga & van de Vijver, 1991; Prieto, 1992; Werner & Campbell, 1970). And, the more advanced measurement methods such as item response theory and covariance structural analysis which have found some use in formally establishing the equivalence of scores obtained from translated tests are not well-known to many persons involved in test tranlations work. Finally the widely used AERA, APA, and NCME Test Standards give only limited attention to the topic (within the framework of item bias analysis). Even in Canada, a bilingual country, and where translating tests from English to French and vice-versa is common, only limited attention is devoted to guidelines for translating tests in the technical standards for tests prepared by the Canadian Psychological Association.

The International Test Commission has organized an international committee of psychologists from the IEA, the European Association of Psychological Assessment, the International Association of Applied Psychology, the International Association of Cross-Cultural Psychology, and the International Union of Psychological Science to begin work on the development of technical guidelines. While it is impossible to predict the outcomes of the committee's deliberations, a review of the test translations literature leads to several preliminary suggestions for guidelines:

1. When it is anticipated or known that a test will be prepared in one language and translated into others, every effort should be made at the item writing stage to use straightforward directions, item stems, and answer choices. Test items with many details are more difficult to translate. Additional suggestions include the repetition of nouns rather than the use of pronouns, avoidance of metaphors, avoidance of the English passive tense (because it's more difficult to translate), and avoidance of hypothetical phrasings or subjunctive mood (Werner & Campbell, 1970). In the mathematics and science areas, for example, conventions about the use of time, money, and units of length, volume, and weight should be agreed upon at the outset. Conventions should insure test

fairness for all examinees. (One convention might be to minimize the number of problems which require units.)

2. Test translators should be chosen for their expertise in the source and target languages <u>and</u> their familiarity with the test content and their experiences in both cultures. (Normally, knowledge of both languages will not be sufficient to produce a satisfactory test translation.) The preferable situation is for test translators to be most familiar with the target language and culture. Knowledge of the principles of writing test items is valuable too.

3. The vocabulary used in two or more versions of the test should be equally familiar to persons in the source and target languages. Test translators can make use of frequency of word use in each language. De-centering can be very helpful, too.

4. Test item formats should be equally familiar to examinees in the source and target languages. When they are not, either provide practice materials to minimize the differences in familiarity or (preferably) use formats that are equally familiar.

5. Cultural differences such as test motivation, vocabulary, experiences (e.g. questions involving time, and measurements of all sorts - length, weight, temperature) should be held to a minimum.

6. The most useful design for establishing the equivalence of two versions of a test requires the source language monolinguals taking the source version and target language monolinguals taking the target version. However, the advantages of this design are lost if statistical techniques are used which require the

questionable assumption of equal ability groups. Use conditional statistical techniques such as IRT or the Mantel-Haenszel procedure whenever possible.

7. When less than ideal data collection designs are used, the shortcomings of the design and the impact of the shortcomings on the validity of the interpretation of results should be clearly noted in the technical manual and along side any interpretations of the results.

8. Whenever possible, both judgmental and statistical methods should be included in a study to determine the equivalence of an original and a translation of the original version of a test.

9. A study of the factorial structures of multiple language versions of a test is valuable in judging the appropriateness of the test translations.

10. Empirical analyses such as comparing the rank order of item difficulties in the two versions of the test is valuable to identify potentially problematic items.

11. Whenever possible, use relatively large examinee samples. Factor analysis, item response models, and the Mantel-Haenszel procedure often require large samples to produce stable results (especially the first two procedures).

12. Documentation on the methods and results of the test translation process should be compiled and organized in a technical report. A good guideline to follow is that the documentation should be prepared to meet the standards of a journal publication.

13. Whenever possible, and certainly for all large scale test translation projects, multiple judgmental and empirical methods

should be used: For example, the process might include (a) training test developers in words, phrases, and concepts to avoid in item writing; (b) evaluation by test translators of the match between the source language and the back-translation of the test; (c) the use of bilingual translators to evaluate the similarity between the source and target language versions of the test; and (d) the collection of data using monolingual examinees taking each version of the test and then subjecting the data to an item bias analysis.

14. Tests should only be administered by persons who are proficient in the language of the test.

Evidence for the suitability of the guidelines above can be found in the psychometric literature. However, the guidelines are only preliminary. A complete, technically sound, and validated list of guidelines should follow soon from the work of the international committee charged with the responsibility of producing guidelines. The work of the international committee to develop technical guidelines for translating tests should be available by spring of 1994, and will be presented at the meeting of the International Association of Applied Psychology in Madrid in July of 1994. In the meantime, it is hoped that the issues, methods, and guidelines discussed in this paper will be useful to those persons who are planning cross-national comparative studies.

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.

Bradburn, N. M., & Gilford, D. M. (Eds.). (1990). A framework and principles for international comparative studies in education. Washington, D.C.: National Academy Press.

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

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

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. The Personnel and Guidance Journal, 56(8), 472-475.

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.

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

De Avila, 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. L. (1986). Assessing the equivalence of measurement of attitudes and aptitudes across heterogeneous subpopulations (unpublished manuscript). Urbana-Champaign, IL: University of Illinois.

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

Ellis, B. B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. Bulletin of the International Test Commission, 18, 33-51.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. Journal of Applied Psychology, 77, 177-184.

Gross, L. J., & Scott, J. W. (1989). Translating a health professional certification test to another language: a pilot analysis. *Evaluation and the Health Professions*, *12*(1), 61-72.

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., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission*, *18*, 3-32.

Hambleton, R. K., & Jones, R. W. (in press). Comparison of empirical and judgmental methods for detecting differential item functioning. *Educational Research Quarterly*.

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

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

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: Lawrence Erlbaum Associates, Inc.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology*, 67, 115-142.

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

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

Jackson, D. N. (1991). Problems in preparing personality test and interest inventories for use in multiple cultures. *Bulletin of the International Test Commission*, *18*, 88-93.

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426.

Joreskog, K. G., & Sorbom, D. (1986). *LISREL VI: User's guide*. Mooresville, IN: Scientific Software, Inc.

Keeves, J. P. (1992). *Learning science in a changing world: Cross-national studies of science achievement, 1970 to 1984*. The Hague, The Netherlands: The International Association for the Evaluation of Educational Achievement.

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.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). Learning mathematics (Report No. 22-CAEP-01). Princetcn, NJ: Educational Testing Service.

Miura, I. T., Okamoto, Y., Kim, C. C., Steere, M., & Fayol, M. (1993). First graders' cognitive representation of number and understanding of place value: Cross-national comparisons -- France, Japan, Korea, Sweden, and the United States. Journal of Educational Psychology, 85(1), 24-30.

Muthen, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dochotomous variables in several groups. Psychometrika, 46, 407-419.

Oakland, T., & Hu, S. (1992). The top 10 tests used with children and youth worldwide. Bulletin ɔf the International Test Commission, 19(1), 99-120.

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

Poortinga, Y., & van de Vijver, F. J. R. (1991). Culture-free measurement in the history of cross-cultural psychology. Bulletin of the International Test Commission, 18, 72-87.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2(3), 255-275.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Arratia, A., Natalicio, L. F. S., & Natalicio, D. S. (1971). Development of the Spanish edition of the state-trait anxiety inventory. Interamerican Journal of Psychology, 5, 145-158.

Swanson, H. L., & Watson, B. L. (1982). Educational and psychological measurement and evaluation. Englewood Cliffs, NJ: Prentice-Hall.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing (pp. 277-308). Boston, MA: Kluwer Academic Publishers.

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. New York: American Museum of Natural History.

Woodcock, R. W. (1985). Woodcock Language Proficiency Battery, Spanish Form: Technical Summary (Assessment Service Bulletin, Number 9). Allen, TX: DLM.

La traduction des tests de rendement : le cas des études internationales.

Ronald K. Hambleton
University of Massachusetts at Amherst, U.S.A.

Pendant fort longtemps, la traduction de tests et de questionnaires conçus dans une langue et une culture afin d'être employés dans d'autres langues et cultures a constitué une pratique répandue. Dans les faits, cependant, tout porte à croire que la qualité des traductions n'est pas constante et que, dans plusieurs cas, les traductions ne sont pas adéquates, ce qui réduit la validité des résultats obtenus au moyen de tels tests et questionnaires. Cette présentation poursuivra trois objectifs précis : (1) envisager quatre difficultés importantes qui se produisent lors de la traduction de tests et voir comment elles peuvent être surmontées; (2) passer en revue les méthodes statistiques et les méthodes fondées sur le jugement d'experts pour établir l'équivalence des scores d'un même test administré en plusieurs langues; et, (3) fournir quelques lignes directrices préliminaires à la fois pour les traducteurs et pour ceux qui réalisent des études d'équivalence.