

Factors Affecting the Quality of the Comparisons of Scores from Adapted Assessments

Linda L. Cook

Educational Testing Service

Introduction

Translating tests into different languages and administering the translated tests to examinees of different cultures is a practice that has a long history in the field of psychological assessment. The early work of both Terman (1916) and Likert (1932) are examples of just how long ago researchers were interested in using instruments developed for one population to assess attributes of a second population that may differ in language, background, and culture.

The practice of translating and adapting tests that were developed for a specific population and then administering the tests to a population that differs in both language and culture, is one that has increased greatly over the past decade. Cook and Schmitt-Cascallar (2005), and Cook (2006) list a number of reasons contributing to the increased interest in test translation and adaptation. One reason is the changing global economy. Businesses are establishing enterprises in multiple countries throughout the world and are employing citizens of the particular country to run these enterprises. This leads to the need to assess the skills of these employees through administration of certification tests in their native languages.

In addition, nations are competing with each other in a global economy and they want to know that their educational systems are producing future employees who will keep their businesses strong in a world market. Consequently, these nations want to compare their students' academic progress to the academic progress of students in other countries. This necessitates the development and adaptation of assessments measuring the same academic skills that are given in multiple languages. International studies of achievement such as the *Third International Mathematics and Science Study (TIMSS)* are an example of this type of work.

It should also be pointed out that many countries, including the United States, are facing an increasing level of diversification of their populations, necessitating academic testing in different native languages. Examples of this are the recent Title III NCLB legislation in the US,

the need to provide academic assessments in both French and English in the Canadian province of Quebec, and the NITE college entrance examinations that are given in multiple languages in Israel.

Each of the purposes for translating and adapting tests that are cited above involves the comparison of scores on tests that are given in different languages to examinees of different cultural backgrounds. The quality of these comparisons depends on the degree of comparability of the scores obtained by the different groups taking the assessments. The focus of this paper is on the factors affecting the comparability of the scores obtained on translated and adapted tests and hence, on the quality of the comparisons of scores that are derived from these tests.

Comparing Scores on Translated and Adapted Tests

Poortinga (1989) points out that comparisons of the abilities of individuals or groups [regardless of the method used for comparison] may be misleading for two reasons. One reason is related to the attribute that is being measured, and he gives as an example the futility of comparing the height of one person to the weight of a second individual. The second reason that Portinga gives for misleading comparisons is related to the scale units used for the comparison; for example, one can not make a direct comparison of the length of two objects if one object is measured in inches and the other in centimeters. These seem like obvious points to make when one is referring to physical attributes such as height, weight, and length that are readily observable. However, the situation immediately becomes more complex when the comparisons are extended to unobservable attributes measured by psychological and educational assessments.

Consider, for example, a test of algebra that contains some word problems. Suppose the test has been constructed in English and scaled using data from a group of English speaking students. The test is then translated into Spanish and administered to a group of Spanish speaking students. If the Spanish speaking students do not score as well on the test as the English speaking students, how do we know whether the differences in scores are because the groups differ in their ability in algebra, or if the score differences are due to the fact that the translation of the algebra word problems into Spanish made the problems some how more difficult for the Spanish speaking examinees?

One possibility is that the test administered in Spanish, requires more reading time than the test administered in English, and consequently the test is more speeded for the Spanish speaking students. Should speed be a factor in the assessment of algebra ability for the Spanish speaking group and not for the English speaking group? In this case, speed would be considered a construct-irrelevant factor affecting the quality of the comparison of scores from the Spanish and English speaking students taking the Algebra test.

A second possibility is that the instructions for the Algebra test might not have been translated clearly and the Spanish speaking examinees might be confused regarding key test-taker strategies, such as whether or not they would be penalized for guessing responses to questions. Again, the quality of the comparison of scores for the two groups of examinees would be affected by construct-irrelevant variance introduced by the translation and adaptation procedures.

The list of reasons for differences between the scores obtained on the algebra test by the Spanish and English speaking groups given above is most certainly not exhaustive, it is only meant to illustrate how difficult it is to avoid construct irrelevant sources of variance in test scores when translating and adapting tests and how the procedures used to translate and adapt tests directly impact the quality of comparisons of scores on these tests.

Procedures for addressing issues of construct irrelevant variance in test scores and consequently promoting an increased level of quality in the comparisons of scores on adapted tests, have been extensively described by a number of researchers. (See Casillas and Robbins, 2005, Geisinger, 1994, and Hambleton, 1993, for thorough discussions of these procedures.) The procedures include translation and back translation of the instrument to be adapted, pilot testing and screening the test items for differential item functioning, field testing and scaling, development of administration procedures, and validation research.

Recognizing the need for validation research on adapted tests is extremely important. This is because, although a meticulous level of attention may be paid to methodological issues when translating and adapting the test, it simply may not be possible to obtain construct equivalence for a test given in multiple languages to groups that differ in language and culture. Consequently, it is important for validation research to be carried out on any translated and

adapted test to ensure that the comparisons and interpretations supported by the test scores are of high quality.

Standard 6.2 of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al, 1985) states that,

“When a test user makes a substantial change in test format, mode of administration, instructions, language, or content, the user should revalidate the use of the test for the changed conditions or have a rationale supporting the claim that additional validation is not necessary or possible.” (p.41)

Many researchers share the opinion of Geisinger, 1994, who stated that, “Such research [validity research] is almost always necessary.”

Threats to the Quality of Comparisons of Scores from Adapted Tests

Hambleton (1996) lists a number of significant threats to the quality of the comparisons of scores obtained on adapted tests. Among these are: lack of construct equivalence; misuse of norming data; poor translation of the test to the target language; failure to adequately link the tests given in the original and target languages; lack of comparability in test administration procedures; and, failure to take into account differences in culture. Geisinger, 1994, adds to Hambleton’s list, issues related to test format.

Earlier in this paper, it was mentioned that, Poortinga (1989) pointed out that the comparisons of scores on adapted tests are impacted by two factors: the attributes that are being measured and the metric used to measure these attributes. Applying Poortinga’s dichotomy, the threats to quality score comparisons attributed to Hambleton and Geisinger can be classified into two groups: those affecting the attributes that are measured; i. e., construct equivalence, translation of the test, test administration procedures, test format, and, cultural differences, and those factors affecting the metric used for comparisons; i. e., linking procedures, and misuse of norming data. The following discussion classifies each of these threats into Poortinga’s dichotomy and shows how each threat is directly related to the quality of the comparisons of scores derived from the translated and adapted tests.

Factors Affecting the Attributes the Tests Are Measuring

Hambleton and Patsula (1998) discuss a set of myths or problems related to adapting tests that they believe undermine most test adaptation initiatives. A prominent myth pointed out by these authors is the myth that constructs are universal and consequently all tests can be translated into other languages and adapted for other cultures. The authors illustrate this point by describing cultural factors, such as speed of response, that may impact scores on intelligence tests. They refer to the fact that the “Western” notion of intelligence places considerable emphasis on speed of response, but that, “In some cultures speed of response is of minor importance as a operating characteristic for life, and members of these cultural groups often score lower on Westernized intelligence tests because of a failure to perform quickly.”

Hulin and Mayer (1986) discussed the issue of construct equivalence across cultures in an article in the *Journal of Applied Psychology*. The authors refer to two conceptions of the role of language in the study of cognitive constructs. One conception, derived from the Sapir-Whorf hypothesis, (Werner and Campbell, 1970), identifies language as the filter between man and the world. According to this hypothesis, language affects experiences and learning to such an extent that cross-language research is impossible.

The second hypothesis is a linguistic position that states that very high fidelity translations from a source to a target language provide a sufficient basis for cross-language and cross-cultural assessments and comparisons. (See, Sireci, 1996, Hambleton, 1996.) Practitioners who are interested in cross-cultural, cross-lingual assessment certainly subscribe to the second conception. But it is important to keep in mind that some theorists believe that it may be impossible to translate tests into different languages and adapt them for populations of different cultures and maintain construct equivalence.

As mentioned above, an essential factor in the equivalence of constructs measured in cross-lingual, cross-cultural studies is the fidelity of the translation of the test to the target language. The difficulty associated with obtaining an adequate translation of a test to the target language is complicated by cultural factors and should not be underestimated. Ahluwalia (1990) discusses the complexity of translating test items into a target language and points out that translations are complicated because of the many cultural differences inherent in languages. She

goes on to say that, “If languages were perfectly parallel, a computer could be used to replace the base language word with its target language equivalent. However, rigid substitution is not possible because language is interwoven with culture.” Ahluwalia illustrates this point by asking the reader to imagine a test question that refers to New York egg creams. Unless the examinee is familiar with this local form of refreshment, it would be impossible to infer that this drink contains neither eggs nor cream. This is a clear example of language that can not be separated from the culture of a particular area of the United States.

Angoff and Cook (1988), point out how cultural factors impact the translation of tests of verbal and mathematical ability adapted from English to Spanish. The researchers were interested in establishing a concordance table that would be used to compare scores on the SAT, administered in English to high school students in the United States, to scores obtained on the Prueba de Aptitud Academica (PAA) administered in Spanish, to high school students in Puerto Rico. As part of the linking process, an effort was made to produce a set of items in English and in Spanish that were, as nearly as possible, equal in meaning in the two languages. These items were translated from one language to the other (either English or Spanish) by a small group of bilingual experts. At a later time, all the items were back translated to their original language and the back translations were compared with the original text. The common item set was then pretested by administering the set to groups of students taking either the PAA or SAT in the appropriate language.

Figures 1 and 2 show plots of the IRT item difficulty parameter estimates (b values) for the verbal and math items respectively. One point that is immediately obvious from an inspection of these two figures is that the plot of the verbal items is much more dispersed than the plot of the math items. The correlation between the b values for the verbal items was .66 and that for the math items was .90. The authors interpreted the greater dispersion of the verbal items as indicating that these items were not measuring “quite the same construct” across the Spanish and English speaking populations. On the other hand, construct equivalence seems to have been more nearly obtained by the mathematical items, although a correlation of .90 does seem to indicate that even in this case there may be some dissimilarities in the construct as measured for the different language/cultural groups.

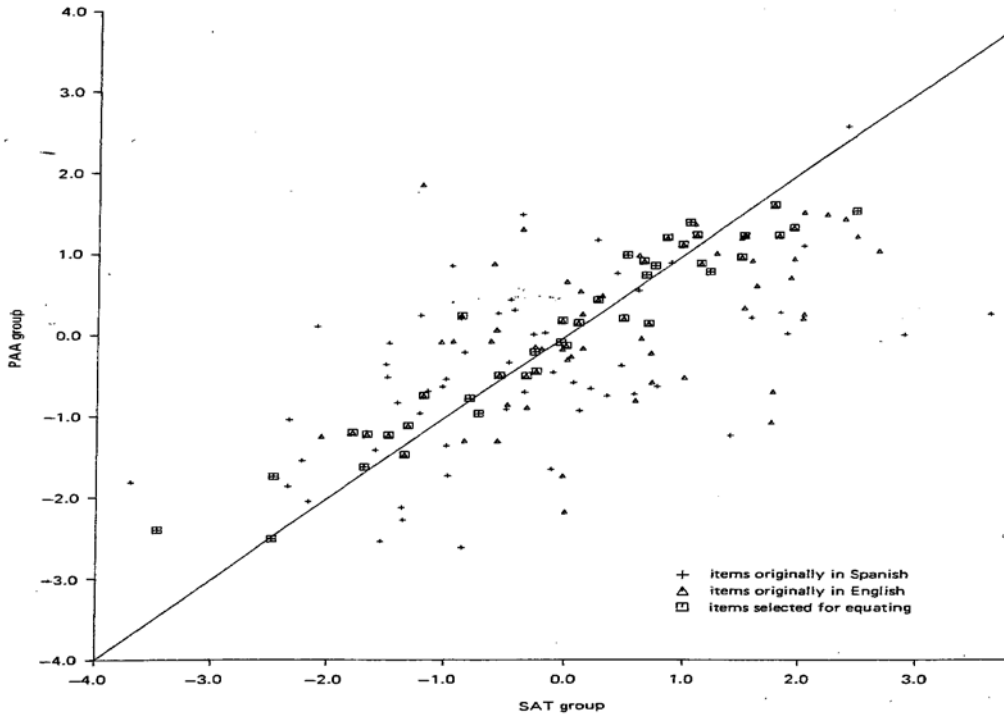


Figure 1. Plot of b 's for pretested verbal items (number of items = 142).

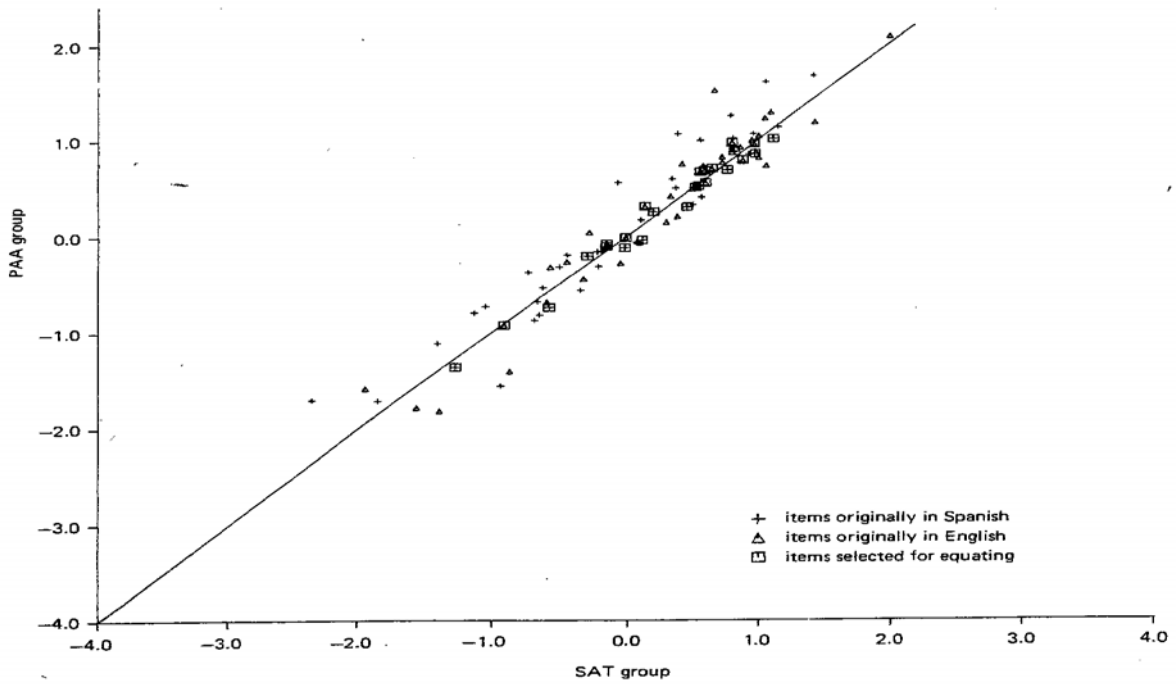


Figure 2. Plot of b 's for pretested mathematical items (number of items = 91).

It is important to note that the dispersion of verbal items shown in Figure 1 could be caused by a number of different factors. Hulin and Mayer (1986) point out some of these factors. They provide, as an illustration, words such as “amigo” and “friend” that may not always be equivalent, but translators may share some rule of thumb that leads them to translate the terms as if they were equivalent. They continue by saying that, “Back translators may produce excellent quality back translations from poorly constructed target language translations by insightful inferences, assumptions, and guesses.” Another reason for unequivalent translations may be the fact that translation in the target language may retain the grammatical form of the original language and may be easy to back translate, but the back translation may not be meaningful to target-language monolinguals.

Because of the problems associated with translating items from the source to the target language, it is essential to guard against differences in constructs, as measured by the adapted and original test, that may have been introduced by the translation process. A considerable amount of research has been carried out on the application of differential item functioning (DIF) procedures for screening translated items in test adaptation applications. (See Elosua and Lopez-Jauregui, 2007, Sireci and Allalouf, 2003, Muniz, Hambleton, and Xing, 2001, for discussions of DIF procedures applied in test adaptation studies.)

As pointed out by Muniz, Hambleton, and Xing, DIF procedures designed to identify differentially performing items across groups that have been matched on ability, have become a routine part of the item analyses carried out for most large scale testing programs. Hence, these procedures are well documented and readily available to practitioners interested in screening items after they have been translated for use in tests adapted for different language and cultural groups.

An additional issue related to the construct equivalence of tests adapted for cross-lingual/cross-cultural use has been pointed out by Geisinger, (1994). The problems pointed out by Geisinger focus on issues related to item and test format. Geisinger, cautions that different cultural or national groups may vary in their levels of sophistication when dealing with different item formats. Geisinger uses a true-false format as an example and states that, “One should not simply use the item format (e.g., true-false) that was used on the instrument in the original culture: rather, those adapting the instrument need to consider the appropriateness of the format

for the [target group].” Geisinger continues by urging that, regardless of the item format used in the adaptation, a sufficient number of exemplary or practice exercises should be used to ensure that the source and target group have equal opportunity to be familiar with the format.

One final factor that may impact the construct equivalence of scores from the source test and the test adapted for the target group is the test administration procedures. Geisinger, points out that it is very important for those adapting tests to develop training programs and materials to instruct users on how to administer, use, and interpret scores from the revised test. Geisinger continues that this is particularly important if the assessment has been translated into a different language. “The language used in the directions for completing the assessment device need to be clear and simple. The materials provided to the test administrator and user, including procedures for scoring the instrument, need to be written so as to minimize potential misunderstandings.” (Geisinger, 1994). It is clear that if, for example, the instructions for a source assessment and the assessment translated into the target language can not be considered comparable, than the tasks that examinees are asked to perform on the two tests, and consequently the constructs measured by the two tests, may not be equivalent.

As mentioned earlier in this paper, it is very difficult to translate and adapt a test constructed for a group that speaks one language and has a particular cultural background, for administration to a second language and culture group. Consequently, studies of construct equivalence of the original and adapted tests, and additional validity evidence for the adapted test, are almost always warranted. Validity studies are typically situation specific, but the essential focus of these studies is on the equivalence of the construct or concept the test is intended to measure. Studies of the validity of an adapted test usually involve determining if the construct or concept measured by the original test remains the same when the test is translated into a different language and administered to a group with a different cultural background.

A classic study carried out by Sireci, Fitzgerald, and Xing (1998) is illustrative of some of the procedures that are available to today’s practitioner for assessing the equivalence of constructs measured by original and adapted tests. The authors evaluated the construct equivalence of different language versions of a certification examination for computer software engineers. Sireci, Fitzgerald, and Xing used three statistical techniques to evaluate construct equivalence: principal components analysis, multidimensional scaling, and confirmatory factor

analysis. The purpose of their research was to determine if the dimensionality of the examination data was consistent across language versions. Consistent results from the dimensionality analyses were interpreted as a basis for an argument for construct equivalence.

The authors found somewhat inconsistent results from application of the three methodologies to tests administered to four language groups. Both the principal components analysis and the multidimensional scaling results indicated that the data were multidimensional, but the confirmatory factor analysis results clearly indicated a unidimensional factor structure across the language groups. The authors concluded that, given the descriptive nature of the principle components and multidimensional scaling analysis, the confirmatory factor analysis results could be interpreted as indicating a single factor structure for all four groups.

Everson, Guerrero, and Laitusis (1998) carried out a confirmatory factor analysis study for the purpose of establishing the construct equivalence of the PAA math test (administered in Spanish) and SAT math test (administered in English) to a bilingual group of test takers. Using exploratory factor analytic techniques, the authors found that the PAA math test appeared to be unidimensional, but the SAT math test appeared to have two factors. After further exploration, the authors concluded that the second factor for the SAT math test was a speed factor possibly related to the English competency of the bilingual sample. The authors carried out a confirmatory factor analysis on the SAT math data after removing 20% of the items at the end of each section. The results of the confirmatory factor analysis indicated that the shortened test was unidimensional.

The studies described above are good examples of the careful analysis that is required to evaluate the construct equivalence of tests adapted for use in cross-lingual/cross-cultural studies. The two studies represent different approaches to the problem of ascertaining construct equivalence. The assessments used for the Sireci, Fitzgerald, and Xing study involved direct translations and adaptations of a single test given to multiple language groups. The Everson, Guerrero, and Laitusis study focused on tests constructed to measure the same construct (math ability) but one test was not a direct translation of the other. In addition the second study differed from the first in that the sample used for the factor analysis was a group of bilingual test takers. The studies are illustrative of the variety of procedures available to the practitioner

interested in adapting tests for administration in different languages to different cultural groups and on evaluating construct equivalence for the tests given to these different groups.

Factors Affecting the Metric Used For Comparisons

As mentioned earlier in this paper, two critical factors that may impact the quality of the comparisons of scores on adapted tests are the proper use of norms data and the establishment of a common metric for the scores that are to be compared.

Assessments are typically developed for use with examinees from a particular linguistic and cultural background. Consequently, information that is developed to aid in the interpretation of scores, such as norming data, is also particular to the language and culture of the source group. Normative information about a test is important because it carries meaning about performance on the test to score users who are interested in making decisions about a specific examinee. For example, college admissions directors can use national and local normative data assembled for the SAT to develop a sense of the level of ability represented by, say, an SAT verbal score of 650. They use this information to estimate how well students who have scores at this level may perform in their particular environment. Geisinger (1994) points out that, “Experienced test users can make highly skilled interpretations on the basis of the proper [normative] test information, especially if they are also able to gather the other information that they need to fine-tune their judgements. But, Geisinger points out that just translating a test and using the same scoring algorithm may not produce scores that can be interpreted with the same meaning associated with scores on the source test. Using norms data based on the source population with scores obtained from the translated and adapted test could be very misleading. Consequently, almost all tests that are adapted into a new language or culture will need to be renormed using data from the appropriate population of test takers so that scores on the new version can be interpreted, studied, and validated for the target population.

In addition to using appropriate score interpretive information for adapted assessments, practitioners interested in comparing scores on these assessments must also be sure that the adapted assessments provide scores in the same metric.

Establishing a common metric for scores obtained on tests given in different languages to examinees with different cultural backgrounds is extremely difficult for many of the reasons pointed out earlier in this paper; particularly those reasons related to the unlikely possibility that the adapted test measures the same construct as the source test once it has been translated into different languages and given to examinees with different cultural backgrounds. The reason this presents a serious problem is that one of the underlying assumptions of most methods that are used to establish a common metric for test scores (scale linking methods) is that the tests that will be linked measure the same, or very similar, constructs.

Ideally, those interested in linking tests that have been translated into different languages and are given to monolingual examinees in their own language, would like to be able to compare the skills and abilities of examinees taking the different tests as though the scores obtained on the assessments were entirely interchangeable (equated). However, this ideal situation is difficult (if not impossible) to obtain because data collected in cross-lingual linking studies does not meet the assumptions made by typical equating models.

Sireci (1996) provides an excellent overview of the technical issues associated with linking tests used in cross-lingual assessments. Sireci begins his review by discussing the fact that some practitioners believe that simply translating a test from one language to another is a sufficient condition for cross-lingual assessment. Sireci points out the fallacy in this line of reasoning by noting that unintended effects of the translation may produce items that differ in difficulty and other characteristics across the different languages. (See also, Geisinger, 1994; Hambleton, 1993; 1996; Olmedo, 1981; and, Prieto, 1992. for a discussion of these issues.)

According to Sireci, methods used to link assessments given in different languages fall into three design categories: 1) separate monolingual group designs; 2) bilingual group designs; and 3) matched monolingual designs. Separate monolingual group designs will necessarily involve some procedure for developing “overlapping items”, whereas the latter two designs have as their central requirement, the development of approximations to overlapping groups of examinees.

Separate Monolingual Group Designs. These designs all involve the administration of tests in original and target languages to their respective language groups and linking the tests

through a set of items that is somehow considered “common” to both language groups. Item response theory (IRT) applications to this type of design have been considered quite promising. IRT models have been used to link tests administered to monolingual groups in several studies. (For examples of these types of studies, see Angoff & Cook, 1988 and O’Brien, 1992).

The major criticism of IRT based monolingual linking studies is that these studies make an untestable assumption about the equivalence of item parameters in the two populations. In other words, the invariant item parameter properties of IRT models are not likely to hold up across the different language samples.

Bilingual Group Designs. Sireci (1996) describes two variants of a bilingual group design. The first design is one in which a single group of bilingual examinees take both language versions of the test in counterbalanced order. Sireci points out that one draw back of this type of design may be practice effects that the counterbalancing may not be able to account for adequately. This is particularly true if the two examinations represent close translations of a single test. A second bilingual design described by Sireci is one in which randomly equivalent bilingual groups each takes a different language version of the tests to be linked. Sireci makes the point that a flaw in this design is the possibility that the random groups are not equivalent.

There are several draw backs to the use of bilingual designs for linking scores on adapted tests. One draw back is that the group may not be truly bilingual and may be stronger in one language than in the other. A second draw back is that the bilingual group may not represent either of the monolingual groups which are the groups of interest in a comparative study. This limitation has serious implications for the generalization of the results of the cross lingual linking study performed using a bilingual group, to monolingual groups.

Matched Monolingual Designs. Given the problems described above with monolingual and bilingual groups designs, the possibility of using a design that matches separate monolingual groups on some of the variables that might affect linking results is quite attractive. However, such designs have rarely been used with success. Matched monolingual designs attempt to bypass the need for common items to assess differences in skills/abilities by using groups for the linking study that are matched on criteria that are relevant to whatever skills or abilities that are assessed by the different language tests.

As Sireci (1996) points out, the effects of matching groups for conventional types of equating designs has been investigated fairly extensively. (See Kolen, 1990; Skaggs, 1990; Cook, Eignor and Schmitt; 1989; Eignor, Stocking, and Cook, 1990; and Livingston, Dorans and Wright, 1990.) The results of these studies are mixed. Livingston, Dorans and Wright suggested that equating may be improved via matching on propensity scores (Rosenbaum and Rubin, 1983), whereas Cook, Eignor and Schmitt advise against such techniques.

A/SAT Linking Studies. A number of important lessons can be learned from an evaluation of the work that has been done over the past 20 years at Educational Testing Service that has focussed on linking scores obtained on the Prueba de Aptitud Academica (PAA) to scores obtained on the Scholastic Assessment Test (SAT). Three studies specifically designed to link scores on the PAA and the SAT have been completed. (See Angoff and Modu, 1973, Angoff and Cook, 1988, and Schmitt, Dorans, Magrina, and Cook, 1998, for discussions of these studies.) Each of the studies conducted attempted to improve on the results of past studies by applying the most current thinking in psychometric theory and the most recent technological developments. Still, even the most recent study carried out by Schmitt, Dorans, Magrina, and Cook, exhibited a number of serious drawbacks. Certainly the experiences gained from the three PAA/SAT linking studies demonstrate how difficult it is to obtain comparable and valid scores on tests that are given to groups that differ in language and culture.

Probably the most significant advance in the Angoff/Modu study (Angoff & Modu, 1973) was the application of the delta plot technique for detecting items in the “common” item equating set that did not behave similarly for the Spanish speaking and English speaking groups. Angoff and Modu used this new technique which initially had been developed for screening items for racial bias (Angoff & Ford, 1973). The procedure involved plotting item difficulty values (deltas) for items administered to the two groups of interest, and deleting those items that fell away from the major axis of the ellipse formed by the plot. Angoff and Modu realized early on that one serious pitfall in cross lingual/cross cultural studies was that in spite of the most meticulous translation and back translation, items that are expected to behave similarly (as “common” items in an anchor test design) must be screened statistically, particularly those items that have a heavy verbal/linguistic component.

The study carried out by Angoff and Cook (1988) built on the design of the earlier study with some methodological and technological improvements. The authors hoped that the use of item response theory (IRT) procedures, to replace the use of the classical test theory procedures that were used in the first study would provide improved results. Indeed, the IRT procedures for detecting differential item functioning (See Lord, 1980), proved to be very powerful procedures for screening common items. The authors were quite confident that by the time they had completed the item screening they were able to construct a “common” test that could be used for linking purposes without risk of advantaging either group of interest. However, the authors began to question the underlying assumptions of the work they were doing. Were the two tests (PAA and SAT) measuring constructs that were similar enough to support the development of a concordance table? What did it mean to use a score on the PAA for a student who spoke only Spanish to estimate a student’s score on the verbal section of the SAT?

As a result of the concerns raised by the authors of the second PAA/SAT linking study, the methodology used for the third study was completely revised. The authors of the third study used regression procedures to develop the relationship between PAA and SAT scores. In the course of the analysis of the data for the third study, they found that the correlations between scores obtained on the PAA verbal and SAT verbal tests (for the bilingual sample used in the study) was only slightly higher than the correlation between scores obtained on the PAA verbal and the PAA math tests. Although there are statistical techniques that can be used to develop concordance tables when tests do not measure the same thing (and, indeed, the work done for the studies just described are excellent examples of this type of work) the question remains, how does one interpret the results of a concordance table developed under these circumstances?

The authors of the third PAA/SAT linking study (Schmitt, Dorans, Magrina, and Cook, 1998) chose to develop prediction equations for predicting SAT scores from PAA scores. The equations took into account not only the verbal or math ability of the examinee as measured by the PAA, but also considered the examinee’s English language ability, as measured by the English Second Language Achievement Test (ESLAT). Although the prediction equations are awkward to use and can not be used readily in the comparison of groups of examinees, they do provide a more accurate answer to the question of how a student who scores at a specific level on the PAA will score on the SAT.

Questions that remain to be explored when considering the linking of the PAA and the SAT are: the description of the similarity or differences between the constructs measured by the PAA and the SAT and how these similarities or differences are impacted by English language ability. In addition, it is important to keep in mind that colleges are not so much interested in predicting SAT scores from a PAA score as they are interested in making valid decisions about how successful students will be if they are admitted to the particular college. The relationship of the scores that was developed in the third PAA/SAT study requires validation by examining the relationship between the predicted scores and performance in college, as measured by freshman grade point average or some other criterion of importance.

Conclusions

The factors underlying the ability to make quality comparisons of scores on two tests that have been adapted for administration to examinees of different language and cultural backgrounds have been explored in this paper. The factors were dichotomized according to Poortinga (1989) and placed into one of two categories: those affecting the attributes measured by the tests; and those affecting the metric used for score comparisons.

Recent research carried out to examine the impact of some of these factors was examined in detail. One clear result of this analysis is that it is very difficult to adapt a test constructed for a particular population to the language and culture of a second population and maintain construct equivalency between the two tests. This result in turn has clear implications for the comparability of the metric of the two measures, since it is well known that true comparability of scores (equating) can only be obtained if two tests measure the same construct (Dorans and Holland, 2000).

Although the issues facing those who wish to adapt tests to multiple languages and cultures and compare scores obtained on these tests are quite daunting, significant advances in methodology have taken place and are continuing to take place every day. Continued high quality research in this area will eventually provide solutions to some of the difficult issues raised in this paper. In the meantime, there is a growing sophistication about the difficulty of carrying out these types of studies and knowledgeable practitioners are beginning to review the results of such studies with the appropriate level of scrutiny.

References

- Ahluwalia, M. S. (1990). Policies for Poverty Alleviation. *Asian Development Review*, 1, 111-132.
- American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing*. Washington, D C: American Psychological Association.
- Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the scholastic Aptitude Test* (College Board Report No. 88-2). New York: College Entrance Examination Board.
- Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 10, 95-106.
- Angoff, W. H., & Modu, C. C. (1973). *Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Research Report No. 3). New York: College Entrance Examination Board.
- Casillas, A., & Robbins, S. B. (2005). Test adaptation and Cross-cultural assessment from a business perspective: Issues and recommendations. *International Journal of Testing*, 5(1), 5-21.
- Cook, L. L. (2006, July). *Practical considerations in linking scores on adapted tests*. Paper presented at the 5th Conference of the International Test Commission, Brussels.
- Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1989, April). *Equating achievement tests using samples matched on ability*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and*

- psychological tests for cross-cultural assessment* (pp. 139-169). Mahwah: Lawrence Erlbaum Associates.
- Dorans, N. J. & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.
- Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of the effects on linear and curvilinear observed-and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education*, 3, 37-55.
- Elosua, P., & Lopez-Jauregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing*, 7(1), 39-52.
- Everson, H. T., Guerrero, A., & Laitusis, V. (1998, April). *Preliminary evidence of construct equivalence of mathematics tests administered across languages: An analysis of findings from the SAT I and the Prueba d Aptitud Academica tests*. Paper presented at the meeting of the American Educational Research Association, San Diego.
- Geisinger, K. F. (1994). Cross-cultural normative assessment: Transition and adaption issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6, 304-312.
- Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, 9, 57-68.
- Hambleton, R. K. (1996, April). *Guidelines for adapting educational and psychological tests*. Paper presented at the annual meeting of the National Council on Educational Measurement, New York.
- Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, 45, 153-171.
- Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology*, 71, 83-94.

- Kolen, M. J. (1990). Does matching in an equating work? A discussion. *Applied Measurement in Education, 3*, 97-104.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 44-53.
- Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73-95.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.
- Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in test translation. *International Journal of Testing, 1*(2), 115-135.
- O'Brien, M. L. (1992). A Rasch approach to scaling issues in testing Hispanics. In K. F. Geisinger (Ed.) *Psychological testing of Hispanics* (pp. 43-54). Washington, DC: American Psychological Association.
- Olmedo, E. E. (1981). Testing linguistic minorities. *American Psychologist, 36*, 1078-1085.
- Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*, 737-756.
- Prieto, A. J. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43*, 1-14.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.
- Schmitt, A. P., Dorans, N. J., Magrina, A., & Cook, L. L. (1998). *Predicting scores on the English Language SAT from the Spanish Language PAA and the Spanish Language English as a Second Language Achievement Test*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

- Sireci, S. G. (1996, April). *Technical issues in linking assessments across languages*. Paper presented at the annual meeting of the National Council on Educational Measurement, New York.
- Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple language and cultures. *Language Testing*, 20(2), 147-165.
- Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). *Adapting credentialing examinations for international uses*. Paper presented at the meeting of the American Educational Research Association, San Diego.
- Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education*, 3, 105-113.
- Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.
- Werner, O., & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Narroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology* (pp. 398-420). New York: American Museum of Natural History Press.