DOCUMENT RESUME

ED 407 433 TM 026 478

AUTHOR Sireci, Stephen G.
TITLE Technical Issues in Linking Assessments across Languages.
PUB DATE Apr 96
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Bilingual Education; Comparative Analysis; *Educational Assessment; *Equated Scores; Item Response Theory; *Language Tests; *Scaling; *Scoring; Second Language Learning; *Test Construction; Testing Programs
IDENTIFIERS *Linking Metrics

ABSTRACT

Test developers continue to struggle with the technical and logistical problems inherent in assessing achievement across different languages. Many testing programs offer separate language versions of a test to evaluate the achievement of examinees in different language groups. However, comparison of individuals who took different language versions of a test are not valid unless the score scales for the different versions are linked or equated. This paper discusses the psychometric problems involved in cross-lingual assessment, reviews linking models that have been proposed to enhance score comparability, and provides suggestions for developing and evaluating a model for linking different language versions of a test. Attempts to link different language versions of a test onto a common scale are classified into three general research design categories: (1) separate monolingual group designs, usually linked through item response theory; (2) bilingual group designs; and (3) matched monolingual group designs. (Contains 4 figures and 47 references.) (Author/SLD)

Technical Issues in Linking Assessments Across Languages[1]

Stephen G. Sireci[2]

University of Massachusetts at Amherst[3]

[1] Paper presented at the annual meeting of the National Council on Measurement in Education as part of the symposium "Cross-Lingual Assessment: Is Close Enough Suficiente?," New York, NY, April 10, 1996.

[2] Over the years my thoughts on this topic have been nurtured through conversations and work with several measurement specialists including Bill Angoff, Giray Berberoglu, Linda Cook, Kurt Geisinger, Ron Hambleton, H. Swaminathan, David Thissen, and Howard Wainer. Their good counsel motivated me to pursue this research. However, the opinions expressed in this paper are my own.

[3] As of this writing, home of the NCAA Final Four Minutemen.

Abstract

Test developers continue to struggle with the technical and logistical problems inherent in assessing achievement across different languages. Many testing programs offer separate language versions of a test to evaluate the achievement of examinees in different language groups. However, comparisons of individuals who took different language versions of a test are not valid unless the score scales for the different versions are linked or equated. This paper discusses the psychometric problems involved in cross-lingual assessment, reviews linking models that have been proposed to enhance score comparability, and provides suggestions for developing and evaluating a model for linking different language versions of a test.

Introduction

Comparing the achievement of students who take different language versions of educational tests is a difficult, if not impossible, task. Such comparisons are troublesome because observed differences in test performance between language groups could be due to either a difference in difficulty between the separate language tests, or to a difference in achievement between the groups. Several methodologies have been applied to the problem of disentangling the "test difference" effect from the "group difference" effect. The objective of these methodologies is to account for the difference in difficulty between separate language versions of a test by transforming the raw test scores from each test onto a common scale. This objective is called "linking" the tests. This paper reviews and evaluates different methodologies for linking tests across languages and provides suggestions for future research in this area.

Purposes of Cross-lingual Assessment

There is a recent increase in the attention paid to cross-lingual assessment. This increase stems in large part from the increasing number of students throughout the U.S. who are not proficient in English, and the desire to compare the educational achievement of students in different countries. In educational and psychological testing, there are numerous examples of the use of tests to compare individuals across languages. Some contemporary examples include:

- comparison of the educational achievement of students in different countries, who receive instruction in different languages (International Association for the Evaluation of Educational Achievement (IEA), 1994; LaPointe, Mead, & Phillips, 1989; Miura, Okamoto, Kim, Steere, & Fayol, 1993),

- evaluation of the cross-cultural generalizability of attitudes or psychological constructs (Ellis, 1989; Hulin, Drasgow, & Komocar, 1982; Hulin & Mayer, 1986; Martin and Berberoglu 1991), and

- evaluation of the academic proficiency of non-English speaking students in the United States with respect to their English-speaking peers (Angoff & Cook, 1988; CTB, 1988; O'Brien, 1992).

Linking different language (DL) tests onto a common scale is also relevant in personnel, licensure, and industrial testing (e.g., Ramos, 1981). Most linking studies in the U.S. have focused on linking tests translated into Spanish to an original English-language version. However, the linking problem is generic across languages. In Israel, for example, the Psychometric Entrance Test, required for entrance into Israeli universities, is linked across six different languages (Beller, 1994).

Test Translation Does Not Signify Equivalence

An intuitive strategy for comparing the educational achievement of individuals who operate in different languages is to translate a test from one language into the other relevant languages. However, it has long been argued that the translation of a test from one language to another does not result in tests that are psychometrically equivalent in both languages (Angoff & Cook, 1988; Geisinger, 1994; Hambleton, 1993; 1994; Olmedo, 1981; Prieto, 1992). Unintended effects of the translation process may produce items that differ in their degree of difficulty across languages. For example, an item might be relatively easy when presented in French, but more difficult when presented in German. Therefore, comparing individuals who took different language versions of a test involves first evaluating the equivalence of the test across languages. Without evaluating translation fidelity, there is no way to determine whether differences observed among the groups are due to "true" group differences, or due to differences between the separate language versions of the test. This is a critical problem for cross-lingual assessment. As Hambleton (1994) pointed out

> The common error is to be rather casual about the test adaptation process, and then interpret the score differences among the samples or populations as if they were real. This mindless disregard of test translation problems and the need to validate instruments in the cultures where they are used has seriously undermined the results from many cross-cultural studies (p. 242).

The recent writings of Hambleton and others regarding problems in cross-lingual assessment have gone a long way in informing the measurement community about the insidious problems in comparing students across languages. A significant contribution to this area of research is the *Guidelines for Adapting Educational and Psychological Tests* forthcoming from the International Test Commission (ITC, in press; largely summarized by Hambleton, 1994). An important point stipulated in the *Guidelines* is that before attempting to link DL tests onto a common scale, it must be demonstrated that the constructs measured by the DL tests are comparable. The focus of this paper is on linking tests presumed to measure equivalent constructs across languages, and so this issue is not addressed. It is further assumed here that the test context and item formats are appropriate for the DL groups. For elaborate discussions of evaluating construct equivalence across languages, see Geisinger (1992, 1994), Hambleton (1993, 1994), Hui and Triandis (1985), Martin and Berberoglu (1991), and Olmedo (1981).

Methods Used to Link Tests Across Languages

Attempts to link different language versions of a test onto a common scale can be classified into three general research design categories: 1) separate monolingual group designs, 2) bilingual group designs, and 3) matched monolingual group designs. These designs are reviewed below. A review of these designs reveals their strengths, limitations, and underlying assumptions.

IRT Linking Using Separate Monolingual Groups

In the separate monolingual group design, source- and target-language versions of a test are separately administered to source- and target-language examinee groups. Items considered to be equivalent across the source- and target-language versions of the test are used to link the DL tests onto a common score scale. The most popular and praised methods for linking via separate monolingual groups use item response theory (IRT) models to calibrate the DL tests onto a common scale. In general, IRT models describe the probability of a particular response to an item by a test taker in terms of characteristics of the item (item parameters) coupled with the relative position of the test taker on the latent variable presumed to be measured by the test. One attractive feature of IRT modeling is that the parameters used to describe the test items are invariant with respect to different samples of test takers who respond to the item. It is this feature, called item parameter invariance, which makes IRT particularly appealing to linking DL tests administered to separate monolingual groups (see Hambleton, Swaminathan, & Rogers, 1991, for a complete description of IRT models).

IRT models have been used in a variety of settings to link DL tests. Educational applications include Angoff and Cook's (1988) linking of the Scholastic Aptitude Test to its Spanish counterpart the Prueba de Aptitud Académica, and O'Brien's (1992) and Woodcock and Muñoz-Sandoval's (1993) linking of English and Spanish language proficiency tests. Examples from industrial testing include the linking of the English and Hebrew, and English and Spanish versions of the Job Descriptive Index (Hulin & Mayer, 1986; and Hulin, Drasgow, & Komocar, 1982). Applications are also found in psychological testing. For example, Ellis (1989) linked English and German intelligence tests, and Martin & Berberoglu (1992) linked English and Turkish versions of a social desirability scale. These applications all used a unidimensional IRT model to calibrate the DL tests; however the particular model employed varied from one study to another.

Although there are variations in the procedures followed in these studies, using IRT to link tests administered to separate monolingual groups typically involves the following steps:

1) The source language (e.g., English language) test is translated into the target language (e.g., Spanish language) via a comprehensive series of adaptation techniques (see Hambleton, 1993; 1994).

2) The source-language test is administered to source-language examinees, and the target-language test is administered to target language examinees.

3) The source- and target-language tests are separately calibrated using an IRT model.

4) A scale transformation procedure (e.g., Stocking & Lord, 1983) is used to place the item parameter estimates for the DL tests onto a common scale. The target-language test item parameters are usually transformed to the source-language test scale.

5) Translated items are evaluated for invariance across the DL tests. IRT-based methods for evaluating differential item functioning (DIF) are typically used to determine item equivalence across languages (see Budgell, Raju, & Quartetti, in press, for a review). The DIF evaluation procedure may be iterative, where items that initially display DIF are eliminated from the subsequent stratifying variable (e.g., "purifying" θ).

6) Items considered invariant across the DL tests are used as "anchor" items to calibrate the tests onto a common scale. Items that are not statistically equivalent across the tests are either deleted or considered unique to the separate language versions. The anchor-item linking procedure could be IRT-based (e.g., concurrent calibration constraining anchor item parameters to be equal), or could be based on a classical anchor-item design.

These general steps do not apply to all studies that used IRT to link DL tests, but are characteristic of the general approach. For example, the Angoff and Cook (1988) study went beyond these general steps by first pre-testing items in English and Spanish populations. This preliminary step allowed them to identify items that appeared statistically equivalent in both populations. The equivalence was re-evaluated with the subsequent calibration sample.

A criticism of the separate monolingual group IRT approach to link DL tests is that the item parameter invariance properties of IRT may not hold over samples derived from DL examinee groups. That is, if the DL groups differ with respect to the proficiency measured, and the calibration procedure does not account for this difference, the parameters for translated items are not directly comparable to their original-language counterparts.

Assumptions Underlying the Monolingual IRT Approach

An evaluation of the assumptions underlying the monolingual IRT approach for linking DL tests reveals the dilemma surrounding item parameter invariance across DL groups. When DL tests are separately calibrated in each language group, the only assumption required for IRT calibration is that the items are measuring a unidimensional construct. However, more restrictive assumptions are invoked when calibrating DL tests onto a common scale. Linking the DL tests requires: construct equivalence across languages, unidimensionality of the pool of DL items, and common items across both tests. This last requirement is the most difficult to realize in practice, and in some cases, it is difficult to determine whether it has been satisfied.

As an illustration of this predicament, consider the monolingual IRT approach outlined above. Without anchor items between the DL tests, it is not possible to link the tests onto a common scale. Concurrent calibration does not form a common scale because differences in proficiency not accounted for by the model would affect the item parameter estimates for the original and translated items. Because only source language examinees take the source language items, the parameters for these items are referenced only to the source language group. Similarly, the target language item parameters are referenced to only the target language examinee group. The sample invariance properties of IRT models may not extend to these DL samples because it is not clear whether the two DL groups represent samples from a single population, or samples from different populations.

The problem of uncertainty of ability differences between groups is easily solved using common anchor items between test forms. Anchor items, by definition, are equivalent in both forms of a test that are to be linked. However, with DL tests, determination of anchor items is problematic. It is clear that translated items cannot be considered equivalent without empirical evidence. But to provide empirical evidence of item invariance across languages, a valid matching criterion is required. The IRT proficiency scale (θ-scale) is a fallible matching criterion because there are no true common items. Scale transformation procedures, such as the Stocking-Lord procedure, do not resolve this dilemma as they require anchor items or some other means for accounting for differences in proficiency between the separate calibration groups.

As an example of the confound between test translation differences and differences between the DL groups, consider two language groups who, on average, differ one-half of a standard deviation unit with respect to the proficiency measured. To make the example more concrete, assume that we are trying to link English- and Spanish-language versions of a multiple-choice science achievement test for junior high school students across English-speaking students in the U.S. and Spanish-speaking students in Costa Rica. Let us assume further that the distribution of science proficiency is the same for the two populations with the exception of the center of the distribution: the Costa Rican distribution centers at θ=.5, while the U.S. distribution centers at θ=0. To link the tests we utilize a monolingual group design using the three-parameter logistic IRT model (Hambleton, et al., 1991). Given this hypothetical "true" difference in science proficiency between these two groups, translated Spanish items with true difficulty parameter up to .5 standard deviation units larger than their English counterparts may appear equivalent if they are calibrated concurrently, or if they are transformed onto a common scale using a procedure that does not account for the difference in group proficiencies.

This predicament is illustrated in Figures 1, 2, and 3. Figure 1 illustrates the hypothetical distribution of science proficiency for these two groups on the hypothetical ("true") English-Spanish scale (θ_T). Figure 2 presents the ICCs for an original and translated item, where the items have different location (difficulty) parameters. Because the true, common, θ-scale accounts for the differences in proficiencies between these two groups, comparing the ICCs illustrates that the item does not function equivalently across the two languages. Obviously, the adaptation of the item from English to Spanish made the item harder. Unfortunately, we do not know θ_T. Figure 3 illustrates how the ICCs would appear if they were scaled concurrently (or transformed onto a common scale) without accounting for the group differences in science proficiency (θ_O is the theta scale estimated from the observed responses). The ICCs in Figure 3 look identical.

Thus the major drawback of the separate monolingual group IRT approach is the inability to separate the DL group proficiency differences from differences due to the DL tests (or items) themselves. Theoretically, the monolingual groups IRT method can be effective only when the equivalence of the anchor items can be defended outside of the IRT calibration model.

Although the IRT approach with monolingual groups involves a potential confound between group proficiency and item nonequivalence across languages, there is some evidence that the procedure works. In the Angoff and Cook (1988) study, the levels of DIF observed across languages were consistent with hypothesized expectations regarding item content and translation difficulty. Items more closely associated with linguistic features displayed DIF more often. Far more verbal items displayed cross-lingual DIF, and the analogy items, which were considered the most context-laden, exhibited the highest level of DIF. Very few mathematics items exhibited DIF. These findings are consistent with what we would expect given a "true" common metric. Thus the example portrayed in Figures 1 through 3, and the associated criticism of the monolingual groups IRT method, may arise only when the item adaptation procedures produce relatively few comparable items. The item adaptation procedures used by Angoff and Cook were comprehensive. It may be that adherence to strict test adaptation guidelines (e.g., Hambleton, 1993; 1994) provides a sufficient number of invariant items for the formation of a common scale for DIF analysis.

Additional problems in calibrating DL tests using separate monolingual groups are non-overlapping portions of the ability distributions for the separate DL groups, and differences between the variance of these distributions. If the DL proficiency distributions overlap only partially, then anchor item equivalence may be possible for only a portion of the θ-distribution for both groups (i.e., only for the interval of overlap). If this problem occurs, then the anchor items used to link the DL tests would not fully represent the distribution of operational items. This is a serious problem because non-representative anchor tests used in anchor-item equating designs have been shown to bias equating results (Cook & Petersen, 1987; Klein & Jarjoura, 1985).

Bilingual Group Designs

One method utilized to separate the effects of group differences across languages from the effects of differences due to the DL tests, is to use a group of examinees who are proficient in both source and target languages (e.g., Boldt, 1969). These bilingual examinees are assumed to be equally proficient in both languages with respect to the proficiency measured. Thus, group differences in proficiency are eliminated, and concurrent calibration is used to calibrate items from the DL tests onto a common scale.

There are three potential variants of the bilingual group design. The most common design is the single-group design where a single group of bilingual students take both language versions of the test (or sets of potential anchor items) in counterbalanced order. This design maximizes language group comparability, but may be affected by a practice effect from taking two tests designed to be identical except for language medium. A second option is to use two randomly equivalent bilingual groups, each of whom takes one language version of the test. This design avoids practice effects, but does not allow for evaluation of the assumption of random equivalence. The third option is to use two randomly equivalent bilingual groups who respond to a mixture of source- and target-language items.

A noteworthy example of the single bilingual group linking design is the method used to link the Spanish Assessment of Basic Education (SABE) to the Comprehensive Test of Basic Skills (CTBS) and the California Achievement Tests (CAT; CTB, 1988). In this study, students who were English-Spanish bilingual responded to pilot sets of Spanish and English anchor items. These items were written to measure the same skills and content areas. The English anchor items were also administered to a monolingual English group and the Spanish anchor items were administered to a monolingual Spanish group. This research design is depicted in Figure 4. The performance of the bilingual group on the pilot anchor items was used to select a set of final anchor items that functioned similarly in both their English and Spanish versions.

The randomly equivalent bilingual groups design was evaluated by Berberoğlu and Sireci (1996). In this study, two randomly equivalent groups of Turkish-English bilingual test takers responded to separate test forms containing English and Turkish polytomously-scored items. Items that were translations of one another appeared on separate test forms, with the exception of two items that were in English on both forms. Using Samejima's (1969) graded response IRT model, they identified items that exhibited "translation DIF," as well as items that were statistically equivalent across the two languages. They concluded that the randomly-equivalent bilingual groups design was an effective procedure for screening items for non-equivalence across languages. They also recommended inclusion of common items across the two forms to evaluate the assumption of randomly equivalent groups.

Although the bilingual group approach directly addresses the problem of disentangling group differences from test differences, it has several major drawbacks. A primary problem is operationally defining "bilingual." It is very difficult to find a group of examinees that are "equally proficient" in two languages (not to mention equally proficient in both languages with respect to the proficiency tested). Bilingual students are not homogeneous with respect to their native language (L1) or second language (L2) proficiency (Baker, 1988; Valdes & Figueroa, 1994). Furthermore, students considered to comprise a bilingual group may differ with respect to the language that is considered to be their native tongue. For example, an English/Spanish bilingual sample may contain primarily students whose first language is English, primarily students whose first language is Spanish, or equal proportions of English and Spanish native speakers.

Another serious problem is the lack of ability of the bilingual sample to represent either group of its monolingual cohorts. A bilingual sample may comprise highly educated students whose bilingualism is accompanied by a multitude of skills above and beyond those possessed by their monolingual cohorts, or it may comprise recently immigrated students who are only marginally proficient in their new language. At best, a sample of bilingual students probably only represents a narrow range of the proficiency distribution of either of their monolingual cohorts. Thus, the results from studies using bilingual test takers suffer from problems of generalizability. The performance of bilingual students may not generalize from one bilingual sample to another, and are not likely to represent either population of monolinguals.

Matched Monolingual Group Designs

The matched monolingual group linking design attempts to control for group differences in proficiency by matching examinees on criteria deemed relevant to the proficiency measured, rather than by accounting for group differences via anchor items. Two approaches can be used: creation of equivalent groups by selecting pairs of examinees in DL groups with similar values on the matching criteria, or using differences between groups on the criteria to account for group differences in the proficiency measured. Caliper matching and matching using propensity scores (Rindskopf, 1986; Rosenbaum & Rubin, 1983) are applicable to this problem. Caliper matching refers to matching on score intervals rather than on exact criterion values. Propensity scores refer to scores that describe "the conditional probability of assignment to a particular treatment given an observed vector of covariates" (Rosenbaum & Rubin, 1983; p.41).

There are not many examples of the matched monolingual group linking design, probably due to the obvious problem of finding relevant and available matching criteria. Tamayo (1990) matched 120 students age 8 to 16 on age, sex, school, grade, and academic achievement (as estimated by their teachers) before evaluating translation differences of the WISC-R vocabulary subtest (32 vocabulary items). Although this approach employed a matched-groups design, it essentially sought out to prove the null hypothesis (i.e., no difference between translated versions of the test) using a relatively small sample, and so the efficacy of this design needs further exploration. An additional disadvantage of the matched group design is that the validity of the matching criteria must be established, and it must be equivalent in both language populations.

Although the matched-groups linking design has not received a great deal of attention in cross-lingual linking studies, matching examinees in DL groups could reduce the effect of group proficiency differences that threaten the validity of the separate monolingual group designs. The effects of matching on equating parallel forms of a test written in the same language have been investigated, but the results are equivocal (Kolen, 1990; Skaggs, 1990). Cook, Eignor, and Schmitt (1989), Eignor, Stocking, and Cook, (1990), and Livingston, Dorans, and Wright (1990) found that matching did not lead to improvement over non-matched designs, while Wright and Dorans (1993) concluded that matching did improve equating results. Wright and Dorans, and Livingston et al., suggested that equating may be improved via matching on propensity scores, but thus far, propensity scores have not been applied to the equating problem. It appears that the idea of matching DL students is intuitively appealing, but is likely to be impracticable.

Comparing the Methodologies: Implications for Future Research

The preceding critique of three methodologies proposed for linking DL tests provides more questions than answers regarding valid cross-lingual assessment. Given the current trend toward cross-national educational comparisons (e.g., Feuer & Fulton, 1994; IEA, 1994), it is clear that ignorance of linguistic factors affecting such comparative studies is unacceptable. It is also clear that accounting for these factors poses formidable challenges for cross-lingual educational researchers.

The review of the literature did not reveal a linking model that completely resolved the problem of linking tests across languages. Of course, it is always easier to point out weaknesses in previous research than it is to provide suggestions for improvement. However, it is not intended here to draw a pessimistic picture of the techniques used for linking tests across languages. Although all methods have their shortcomings, they go far beyond the assumption that scores derived from DL tests are directly comparable. These state-of-the-art techniques represent considerable progress from the earlier days of cross-cultural research where differences in test content across languages were not even considered as potential confounds affecting observed group differences (Brislin, 1970; Hambleton, 1994; Prieto, 1992). Rather, the designs reviewed in this paper are far superior methods for promoting score comparability across DL tests than are methods that employ translation only, or that use "expert" judgment to certify score equivalence.

Obviously, the most obstinate problem in linking DL tests is accounting for the differences in proficiency between the DL groups. Procedures that use anchor items to account for group differences suffer from a serious theoretical flaw; items that are translations of one another cannot be assumed to be equivalent, and so they are poor anchor items. Similarly, IRT methods used to evaluate translation DIF (e.g., Budgell et al., in press) provide no way of determining the effect of unknown group differences on the estimated item parameters. Thus, future research should focus on identifying items that are truly invariant across languages, whose invariance can be established independently of a particular calibration model.

Non-verbal items, or items minimally associated with linguistic content, are a potential solution to this problem. The equivalence of such items across languages is likely to be defendable irrespective of statistical evaluation. In educational testing, it is extremely difficult to envision items free of linguistic elements. However, the observational techniques used in some psychological assessments, such Ainsworth's "strange situation" (Ainsworth, Blehar, Waters, & Wall, 1978) assessment of mother/infant interaction, are truly language free, and have been used successfully to evaluate psychological constructs across DL groups (Shelley-Sireci, Fracasso, Busch-Rossnagel, & Lamb, 1995). Perhaps emerging performance-based educational assessments will yield items that minimize linguistic effects. For example, science tests could ask examinees to identify elements in the periodic table with specific properties (e.g., 3 electrons), choose the chemicals required to neutralize an acid, or complete an unfinished drawing illustrating the flow of magnetic forces. Such items could then be used to set a common metric for evaluating translation DIF.

A promising area of future research is evaluating the effects of increased rigor in the test translation process. The few studies that have linked tests across languages provide provocative preliminary evidence that rigorous translation procedures facilitate item equivalence across languages (Angoff & Cook, 1988). Adherence to the test adaptation guidelines currently promoted by the ITC (Hambleton, 1994) should reduce the likelihood of introducing biasing factors into the translation process.

Innovative research designs incorporating subgroups of bilingual test takers may also address some of the shortcomings of approaches using monolingual groups. For example, Berberoğlu and Sireci (1996) found that when bilingual examinees were presented with items that were more ambiguous in L2 than in L1, students were more hesitant to endorse extreme positions on the Likert scale associated with the L2 versions of the items. They concluded that bilingual test takers could not be used to link DL tests, but could be used to identify items that were not equivalent given a bilingual sample. What is missing from the literature is a comprehensive study that uses several types of bilingual groups in conjunction with source and target language groups. Future research should evaluate different types of bilingual test takers, who vary in their degree of facility with both languages and who are counterbalanced according to native language. Linguists critical of testing bilingual and ESL students hypothesize that monolingual tests prevent bilingual students from demonstrating knowledge that is best communicated in the non-test language. Future research should test this hypothesis. For example, randomly equivalent groups of bilingual students could be assigned mononlingual or bilingual versions of a test to evaluate whether restricting their responses to the L1 or L2 language impedes their performance. Further research on the test performance of diverse groups of bilingual students is likely to illuminate problems and solutions relevant to cross-lingual assessment.

Future research should also explore matching DL monolingual examinees to tease out the effects of language-group proficiency differences from differences due to the test translation process. Matching via propensity scores is theoretically appealing, but has not been evaluated with respect to linking DL tests. As with the bilingual group design, matching DL groups will probably not result in a defendable linking design in its own right, but may be useful for supplementing designs using separate monolingual groups.

An emerging area of research that is also relevant the linking problem is multidimensional IRT models (e.g., Ackerman, 1994). If separate dimensions can be identified for source or target language proficiency, and the proficiency purportedly measured by the test, then the latter dimension can be used as a "purified" matching criterion for evaluating DIF among original and translated items.

Nonlinguistic anchor items, stricter test adaptation procedures, bilingual group research designs, matching strategies, and multidimensional IRT models are promising possibilities for enhancing the score comparability of DL tests. Empirical research is needed to determine their utility. In addition to the technical problems of linking DL tests, questions of construct and predictive validity must also be evaluated further (Anastasi, 1992; Geisinger, 1992; 1994; Hambleton, 1993; 1994). Nevertheless, when test score-based inferences focus on comparing the proficiencies of DL examinees, adjustment for differences due to the measurement procedure (i.e., linking) is requisite. Ignoring the effects of multiple languages in a global society severely limits the validity of contemporary educational research. Realizing the limitations of cross-lingual assessments is a necessary first step towards resolving these difficult measurement problems.

References

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 7, 255-278.

Ainsworth, M., Blehar, M., Waters, E., & Wall, S. (1978) . Patterns of attachment. Hillsdale, NJ: Erlbaum.

Anastasi, A. Introductory remarks. In K.F. Geisinger (Ed.) Psychological Testing of Hispanics (pp. 1-7). Washington, DC: American Psychological Association.

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.

Baker, C. (1988). Normative testing and bilingual populations. Journal of Multilingual and Multicultural Development, 9, 399-409.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli Universities. Educational Measurement: Issues and Practice, 13, 12-20.

Berberoğlu, G. , & Sireci, S.G. (1996). Evaluating Translation Fidelity Using Bilingual Examinees. Laboratory of Psychometric and Evaluative Research Report No. 285. Amherst, MA: University of Massachusetts, School of Education.

Boldt, R.F. (1969). Concurrent validity of the PAA and SAT for bilingual Dade School County high school volunteers. College Entrance Examination Board Research and Development Report 68-69, No. 3, Princeton, NJ: Educational Testing Service.

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

Budgell, G.R., Raju, N.S., & Quartetti, D.A. (in press). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement.

CTB (1988). Spanish assessment of basic education: Technical report. Monterey, CA: McGraw Hill.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1989). 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. & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

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.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74, 912-920.

Feuer, M.J., & Fulton, K (1994). Educational testing abroad and lessons for the United States. Educational Measurement: Issues and Practice, 13, 31-39.

Geisinger, K.F. (1992). Fairness and selected psychometric issues in the psychological testing of Hispanics. In K.F. Geisinger (Ed.) Psychological Testing of Hispanics (pp. 17-42). Washington, DC: American Psychological Association.

Geisinger, K.F. (1994). Cross-cultural normative assessment: translation and adaptation 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. (1994). Guidelines for adapting educational and psychological tests: a progress report. European Journal of Psychological Assessment, 10, 229-244.

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

Hui, C.H., & Triandis, H.C., (1985). Measurement in cross-cultural psychology: a review and comparison of studies. Journal of Cross-Cultural Psychology, 16, 131-152.

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

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

International Association for the Evaluation of Educational Achievement (1994). TIMSS main study manuals: population 1 and 2. Hamburg: Author.

International Test Commission (in press). Guidelines for adapting test instruments and establishing score equivalence. European Journal of Psychological Assessment.

Klein, L.W., & Jarjoura, D. (1985). Effect of number of common items in common-item equating with nonrandom groups. Journal of Educational Measurement, 22, 197-206.

Kolen, M. J. (1990). Does matching in an equating work? A discussion. Applied Measurement in Education, 3, 97-104.

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

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.

Martin, M.R., & Berberoglu, G. (1991). Initial efforts in construct validation for the Turkish Marlowe-Crowne Social Desirability Scale. In B. Thompson (Ed.), Advances in educational research: substantive findings, methodological developments (pp. 25-36). Greenwich, CT: JAI Press.

Martin, M.R., & Berberoglu, G. (1992). Further construct validation of the Turkish Marlow-Crowne Social Desirability Scale using the Rasch model. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX, February 1.

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 comparison -- France, Japan, Korea, Sweden, and the United States. Journal of Educational Psychology, 85, 24-30.

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

Prieto, A. J. (1992). A method for translation of instruments to other languages. Adult Education Quarterly, 43, 1-14.

Rindskopf, D. (1986). New developments in selection modeling for quasi-experimentation. In W. M .K. Trochim (Ed.) Advances in quasi-experiential design and analysis, no. 31. San Francisco, CA: Jossey-Bass.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement No. 17, 34 (4, Pt. 2).

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. Applied Measurement in Education, 3, 53-71.

Shelley-Sireci, L.M., Fracasso, M.P., Busch-Rossnagel, & Lamb, M.E. (1995). Mother-infant social and instrumental interaction in culturally diverse populations. Paper presented at the annual meeting of the American Psychological Society, July, New York, NY.

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.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Tamayo, J.M. (1990). A validated translation into Spanish of the WISC-R vocabulary subtest words. Educational and Psychological Measurement, 50, 915-921.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. European Journal of Psychological Assessment, 3, 1-16.

Wright, N. K., & Dorans, N. J. (1993). Using the selection variable for matching or equating. (Research Rep. No. RR-93-4). Princeton, NJ: Educational Testing Service.

Figure 1: Hypothetical Proficiency Distributions
("True" Common Scale)

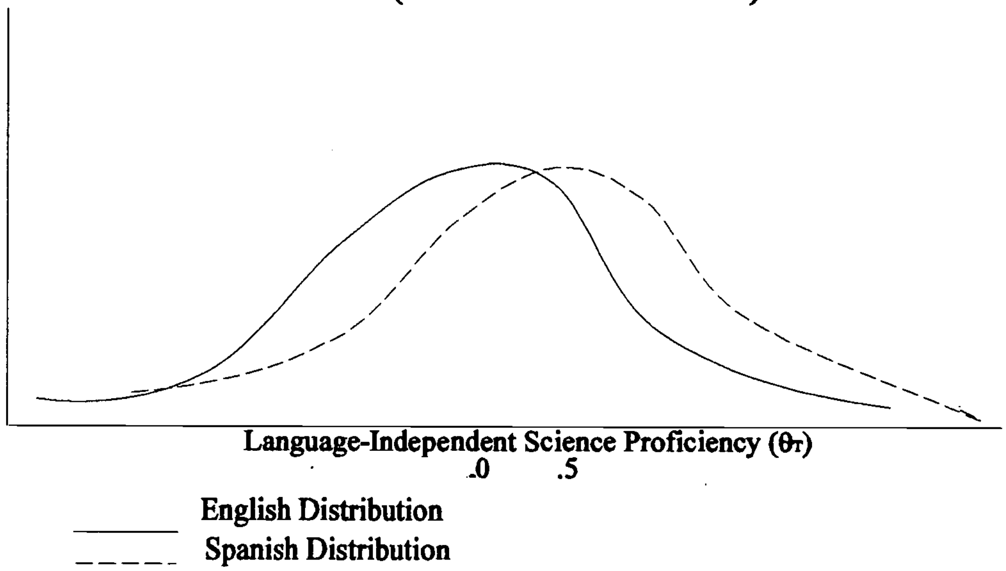


Figure 2: Original & Translated Item On Hypothetical Common Scale

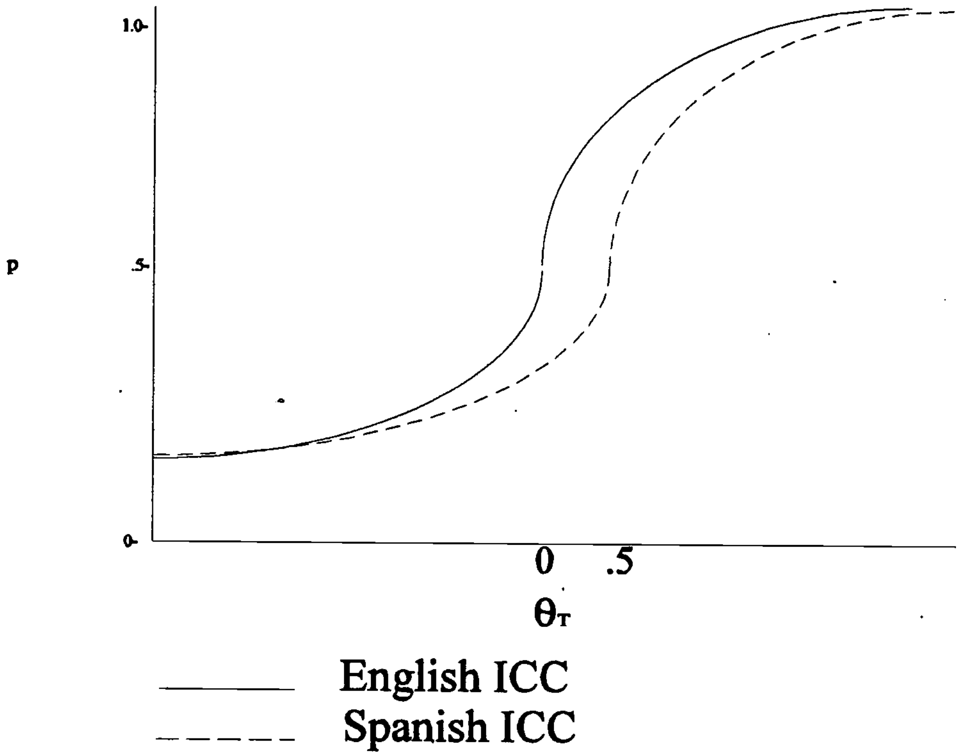


Figure 3: Concurrently-Calibrated ICCs

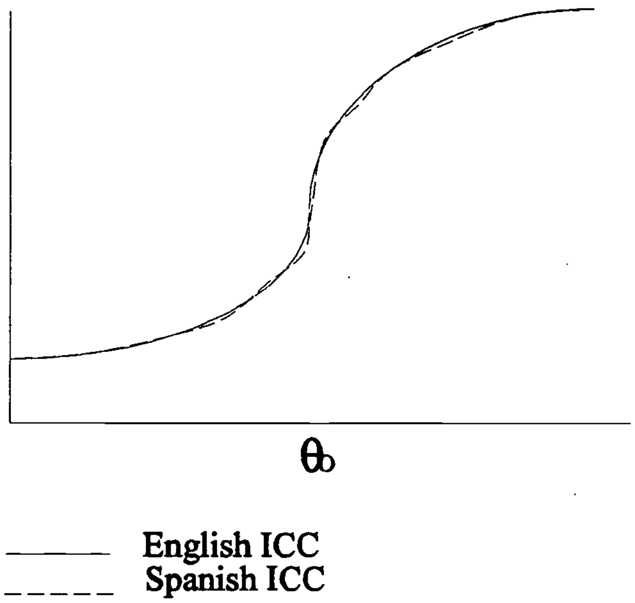
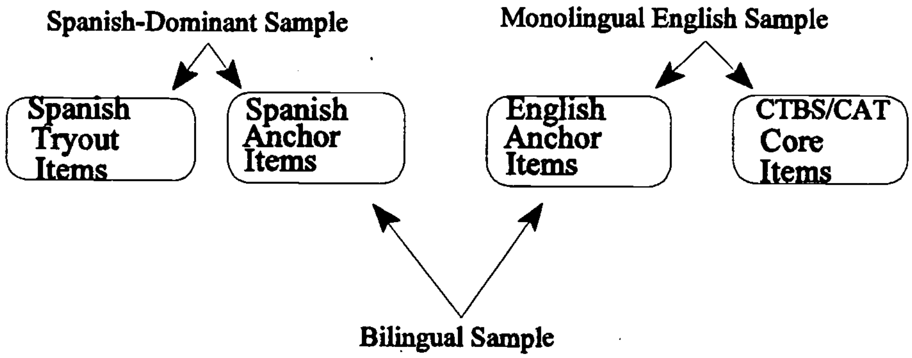


Figure 4: Schematic of SABE Research Design
(from CTB, 1988, p. 6)



TM026478

NCME Annual Meeting, April 9-11, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Technical Issues in Linking Assessments Across Languages	
Author(s): Stephen G. Sireci	
Corporate Source: School of Education University of Massachusetts Amherst	Publication Date: April, 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ ← **Sample sticker to be affixed to document**

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

_____ *Sample* _____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document → □

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

_____ *Sample* _____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature:	Position: Assistant Professor
Printed Name: Stephen G. Sireci	Organization: University of Massachusetts Amherst
Address: School of Education University of Massachusetts Amherst, Massachusetts 01003	Telephone Number: (413) 545-0564 Date: January 27, 1997

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

March 12, 1996

Dear NCME Presenter,

Congratulations on being a presenter at NCME[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the NCME Conference. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (23)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

NCME 1996/ERIC Acquisitions
O'Boyle Hall, Room 210
The Catholic University of America
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the NCME web page (http://www.assessment.iupui.edu/ncme/ncme.html). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an NCME chair or discussant, please save this form for future use.