ED 428 119 TM 029 590

AUTHOR Sireci, Stephen G.; Swaminathan, Hariharan

TITLE Evaluating Translation Equivalence: So What's the Big Dif?

PUB DATE 1996-10-00

NOTE 15p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 1996).

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Cross Cultural Studies; Foreign Countries; International Education; *Item Bias; Language Proficiency; *Second Languages; *Test Format; Test Use; *Translation

IDENTIFIERS *Dimensionality (Tests); Item Bias Detection

ABSTRACT

Procedures for evaluating differential item functioning (DIF) are commonly used to investigate the statistical equivalence of items that are translated from one language to another. However, the methodology developed for detecting DIF is designed to evaluate the functioning of the same items administered to two groups. In evaluating the differential functioning of dual language versions of a test item (translation DIF), the items being compared (i.e., an original item and its translated version) are not the same. Thus, studies of translation DIF may not fulfill the requirements of currently available DIF detection procedures. This paper discusses the complex issues involved in evaluating translation DIF. An important, but often overlooked, issue is the dimensionality of the construct measured across the two languages. It is concluded that the dimensionality issues must be addressed first in studies of translation DIF. The development of an adequate research design is another important issue in studies of translation DIF. The design must be able to control for extraneous language proficiency effects. Some suggestions and examples of such research designs are proposed. (Contains 29 references.) (Author/SLD)

Evaluating Translation Equivalence: So What's the Big Dif?[1]

Stephen G. Sireci and Hariharan Swaminathan

University of Massachusetts, Amherst

[1] Paper presented at the annual meeting of the Northeastern Educational Research Association as part of the symposium "Issues in the Detection of Differential Item Functioning," Ellenville, NY, October, 1996

Evaluating Translation Equivalence: So What's the Big Dif?

Stephen G. Sireci and Hariharan Swaminathan

University of Massachusetts, Amherst

Abstract

Procedures for evaluating differential item functioning (DIF) are commonly used to investigate the statistical equivalence of items that are translated from one language to another. However, the methodology developed for detecting DIF is designed to evaluate the functioning of the <u>same</u> items administered to two groups. In evaluating the differential functioning of dual language versions of a test item (translation DIF), the items being compared (i.e., an original item and its translated version) are <u>not</u> the same. Thus, studies of translation DIF may not fulfill the requirements of currently available DIF detection procedures. This paper discusses the complex issues involved in evaluating translation DIF. An important, but often overlooked issue is the dimensionality of the construct measured across the two languages. It is concluded that the dimensionality issue must be addressed first in studies of translation DIF. The development of an adequate research design is another important issue in studies of translation DIF. The design must be able to evaluate the dimensionality of the construct measured across the two languages, and must be able to control for extraneous language proficiency effects. Some suggestions and examples of such research designs are proposed.

Statement of the Problem

The practice of translating a test from one language to another is almost as old as the practice of testing itself. However, for years psychometricians have strongly warned against assuming the equivalence across languages of translated versions of a test (e.g., Likert, 1932; Geisinger, 1994; Hambleton, 1993; Olmedo, 1981). Nevertheless, researchers have all too often drawn strong conclusions about educational or psychological differences between individuals who operate in different languages without considering the potential effects of different language versions of the measuring instrument (Hambleton, 1994).

In the past two decades several technological developments aimed towards evaluating the equivalence of dual-language versions of tests have emerged (Sireci, in press). Most of these developments emerged from the *differential item functioning* (DIF) methodology, which is designed to evaluate the stability of an item's functioning across examinees in different groups who are at the same proficiency level of the construct measured (Budgell, Raju, & Quartetti,1995; Green, 1994; Holland & Wainer, 1989; Swaminathan, 1994). DIF procedures are typically employed to evaluates how items function across different groups of individuals who operate in the same language. For example, a question that is often addressed is: does an item function equivalently for Black and White examinees? Thus, currently available DIF detection procedures are designed to evaluate the functioning of the <u>same</u> items administered to two groups. In evaluating the differential functioning of dual language versions of a test item (translation DIF), the items being compared (i.e., an original item and its translated version) are <u>not</u> necessarily the same. Therefore, popular DIF detection procedures, such as IRT-based, standardization, Mantel-Haenszel , and logistic regression methods, should not be used blindly in the study of translation DIF.

Obviously, there is a "big Dif" between DIF studies conducted using two same-language groups of examinees on the same group of items, and those conducted using two different-language groups of examinees on different-language versions of the items. The purpose of this paper is to present the complex issues involved in evaluating the differential functioning between items that are translated from one language to another. Only after considering these complex issues can appropriate methodologies for evaluating translation DIF be proposed.

Issue 1: Confounding of Examinee Group Effects and Test Language Effects

The dual-language DIF situation involves teasing out the effects of two factors: language group differences and item language differences. It is almost impossible to successfully isolate the effects of both factors. Without common items to evaluate group differences fair comparisons across groups cannot be made. Without accounting for group differences, differences between original and translated items cannot be made.

Research on DIF methodologies has revealed that when large differences in proficiency exist between the two groups being compared (i.e., when substantial *impact* is present) incorrect conclusions about DIF can be drawn (Green, 1994; Swaminathan, 1994). This problem stems in part from the inability to "match" examinees in the different groups who are located in areas where the proficiency distributions do not overlap. This inability to match examinees poses a very serious problem in cross-lingual DIF analyses for three reasons. First, groups who operate in different languages, for example individuals from different countries, are likely to be different with respect to the attribute measured. Secondly, unlike the monolingual DIF situation, there is no direct way of measuring impact (the overall difference between the groups). Thus, cross-lingual DIF analyses, whether IRT-based or other, typically have no means of incorporating group differences into the model (Sireci, in press). A third problem is that the dimensionality of the proficiency distributions for each language group may be different; in the worst case, the proficiency distributions may represent different constructs altogether.

Given a unidimensional assessment (we consider multidimensionality later), how does one separate test translation effects from language group effects when evaluating translation DIF? A solution can be derived if one of two assumptions are made: 1) assume the items are equivalent across the two languages and then look for group differences; or 2) assume the groups are equivalent and then look for item translation differences. The sample invariance feature of item parameters estimated using item response theory (IRT) make choosing the first assumption attractive. For if a sufficient number of items are equivalent across languages, and the unidimensionality requirement holds, the IRT theta scale can be used as both a conditioning variable for evaluating DIF, and for determining differences across the language groups. It is perhaps for this reason that several applications of IRT DIF detection procedures have been applied to the "translation DIF" problem (Angoff & Cook, 1988; Ellis, 1989, Hulin & Mayer, 1986). However, the asumption of unidimensionality and translation equivalence for the majority of items may result in invalid interpretations regarding translation equivalence if one or both of these assumptions do not hold. Therefore, these assumptions must be empirically tested when applying unidimensional DIF detection procedures, especially IRT based procedures, to the translation DIF problem.

Issue 2: Dimensionality of the Construct Measured

There are several important dimensionality issues that need to be considered when studying translation DIF. First, if unidimensional DIF detection procedures are to be used, the dimensionality of the instrumnt in each language must be assessed. Second, it must be demonstrated that the same unidimensional construct is measured by both language versions of the test. Additionally, if the construct is determined to be multidimensional, the common dimensions across languages must be identified. Therefore, to evaluate translation DIF, the dimensionality of the construct measured, and the degree of equivalence of the construct across languages, must be understood.

Principal components analysis, factor analysis, and structural equations modeling are procedures that are applicable for evaluating test dimensionality. Structural equations modeling is the most comprehensive procedure for evaluating construct equivalence across languages because it can handle several groups and measurement variables within a single model, and provide statistical tests of relative model fit. Unfortunately, these procedures are based on the assumption of linear relationships between the observed variables and the the underlying constructs or latent variables. When the items in the measuring instrument are dichotomoulsy or polytomously scored, the assumption of linearity is violated. Unfortunately, non-linear structural equations models, that would be appropriate for dichotomously-scored test data, are not currently available. However, some advances have been made on non-linear factor models that are appropriate for the dimensional analysis of dichotomous data (Mcdonald, 1968), Bock & Aitken (1981) and these procedures have been implemented in the computer packages NOHARM (Fraser, 1983) and Testfact (Bock, Gibbons, & Muraki, 1988).

Aside from the issue of how best to evaluate dimensionality across languages, there is the complex issue of how to gather the data that would be appropriate for examining translation equivalence. Thus, a third important issue in evaluating translation DIF revolves around the choice of an appropriate resarch design.

Issue 3: Research Design Options in Evaluating Translation Equivalence

There are at least threel research designs that could be used to evaluate translation equivalence. These designs can be classified as one-group, two-group, and four-group designs. These designs all use bilingual subjects to assess the equivalence of the translation.

We need a description of each of these designs together with their strngths and weaknessess. We could shorten the literature review provided here

The Single Group Design

One design that is recommended is to use bilingual examinees to take the different language tests. This design attempts to remove the test effect/group effect confounding by using one or more groups proficient in both languages. A single-group bilingual design could be used where all examinees take both language versions of the test, or random assignment could be used to gather data on each language using a separate, but randomly equivalent bilingual group. Details on some bilingual research designs are presented below. Previous research discussed the role of bilinguals in evaluating translation DIF (Sireci, in press; 1996). However, bilinguals can also be used to evaluate construct dimensionality and comparability across languages.

As an example, suppose we use a single bilingual group and administer both language versions of a test. A score can be calculated for each examinee on both tests (e.g., separate IRT theta estimates) and the correlation between these two scores can be determined. If the two sets of scores are highly correlated, evidence that the same construct is measured in both languages is provided. One drawback to this design is that a testing effect may be present. Because examinees take both test forms, responding to the items on one form may influence responses to the translated versions of the items on the other form. Using random assignment of bilinguals to test forms would remove the practice effect, but would also prevent calculation of the correlation between scores, unless strong matching techniques were used (i.e., matching two examinees on proficiency in each language and on the proficiency measured by the test, and then assigning each randomly to one test form). Both design strategies could also be used to evaluate test dimensionality. Including items of both language types on a single test form (see below) may be particularly useful in studying dimensionality using bilinguals. If dimensions are identified on which only same-language items exhibit substantive loadings, the degree to which the language effect contributes to multidimensionality could be determined.

The logic in using bilinguals to evaluate different language tests is that by using a single group of examinees, language group differences are eliminated, and observed differences in test or item performance can be attributed to the linguistic differences between the tests or items. There are several applications of research designs using bilinguals which have contributed greatly to understanding differences across dual language versions of tests (e.g., Boldt, 1969; CTB, 1988; Berberoglu & Sireci, 1996). However, although the use of a single group usually eliminates group differences in most research designs, there are some deficiencies in this logic when applied to the cross-lingual assessment situation. The most conspicuous problem is the implicit assumption that bilinguals are equally proficient in both languages. For example, if a group of bilinguals performs differently on the "language A" and "language B" versions of an item, attributing this difference to a faulty adaptation assumes that the bilinguals would perform the same on both language versions of the item if the adaptation was adequate. However, a bilingual examinee may be stronger in one language than the other. Therefore, a plausible rival hypothesis is that bilinguals perform better on items administered in their stronger language, even when the two versions of the item are truly equivalent.

A second flaw in this logic is that it describes bilinguals as a single, homogeneous group of test takers; when in reality, a bilingual group of test takers comprises examinees with very different backgrounds, proficiencies, and linguistic skills (Baker, 1988; Valdés & Figueroa, 1994). For example, in the U.S., a group of English-Spanish bilinguals could include people whose first language is English and who learned Spanish in high school, Spanish-speaking immigrants (from a wide variety of countries) who recently learned to speak English, and second-generation immigrants who learned English as a second language in primary school. Therefore, the assumption that bilinguals represent a single "kind" of test taker is unreasonable.

A more subtle, but serious problem in using bilinguals to evaluate tests is the questionable comparability of bilinguals and monolinguals (Hambleton & Kanjee, 1995). This problem has been termed the representation problem (Sireci, in press). For example, in educational testing, bilinguals are likely to be very different from their monolingual cohorts. Bilinguals who are highly proficient in two languages may be representative of only the highest achieving students in either monolingual group. Conversely, bilinguals who are marginally proficient in one or both languages may represent only the lowest achieving students of one of the monolingual groups. In any event, the distribution of proficiency in a bilingual sample is likely to be very different from the corresponding distributions of their monolingual cohorts.

Berberoglu and Sireci (1996) acknowledged the problems in using bilinguals to evaluate translation DIF, but argued that such analyses can be a useful first step in evaluating item functioning across languages. Using a group of Turkish-English bilingual examinees, they applied Samejima's (1969) graded response IRT model to a set of polytomously-scored teacher evaluation items, and used IRT likelihood ratio DIF detection procedures (Thissen, Steinberg, & Wainer, 1988) to evaluate the equivalence of the Turkish and English versions of the items. They concluded that three of the seven items studied exhibited translation DIF. Using these results, they were able to identify factors that may explain why these items functioned differentially across the two languages, while the other items did not. Nevertheless, the authors concluded that their procedure was not appropriate for determining whether any of the items would exhibit DIF across monolingual groups of Turkish and English examinees.

Berberoglu, Sireci, and Hambleton (1997) extended the Berberoglu and Sireci study by including DIF analyses based on both bilingual (Turkish-English bilinguals) and monolingual (English-only) examinees. Following the recommendations of the earlier study, the bilingual group was used to identify anchor items for subsequent DIF analyses using separate monolingual groups (actually, the data from the Turkish-English bilingual group also served as the Turkish monolingual group). Item equivalence was evaluated both with and without the use of anchor items selected based on the bilingual analyses. The rationale was that items that did not display DIF using bilinguals could be considered appropriate for calibrating the remaining items onto a common scale. The effect of using this anchoring was explicitly evaluated. Berberoglu et al. hypothesized that:

1) The DIF items identified using the bilingual group would also be identified as DIF items when the analyses using the "monolingual groups" included anchor items to form a "common" scale.

2) Differences in the detection of translation DIF items would occur depending on whether anchor items were used in the monolingual analyses. It was assumed the analyses using anchor items would identify "true" DIF items, whereas the DIF results from the analyses not using anchor items would be questionable..

Thus, the Berberoglu et al. study predicted that using anchor items selected from the bilingual group analyses would lead to substantive improvement in the detection of translation DIF.

The results of Berberoglu et al. were contrary to their expectations. The two items that exhibited DIF using bilinguals did not exhibit DIF in the anchor item monolingual comparisons. Moreover, the use of anchor items in the monolingual analyses had no effect on DIF detection. The same items were identified as DIF items in both the anchor item and non-anchor item analyses.

The authors speculated that a difference in the English proficiency of the Turkish-English bilinguals and the monolingual-English examinees could account for the results. If the Turkish-English bilinguals misinterpreted the English version of an item, it may exhibit DIF within that population due to a problem of limited English proficiency. When interpreted appropriately in English, the item may actually be equivalent to its Turkish counterpart. Thus, using bilinguals to evaluate translation DIF appears to confound language proficiency differences with translation DIF. This potential confound is especially problematic when only one group of bilinguals are used. Had English-Turkish bilinguals been used (i.e., native English speakers who were proficient in Turkish), more accurate results may have been obtained (Sireci, 1996).

Berberoglu et al., also indicated other limitations of the study that may have affected the results. Most notable is that some changes were made to the English version of the test between the time the bilinguals took the test and the time the English monolinguals took the test. Thus, these data do not provide a clear answer regarding the utility of bilinguals for selecting anchor items. Presently, the only conclusion that can be derived from these studies is that using bilinguals does not solve the translation DIF problem.

The Two-Group Design

Single-group bilingual designs do not address language proficiency differences among bilinguals. One potential way to address this shortcoming is to use a more than one group of bilinguals. The simplest design involves two randomly equivalent groups of bilinguals. These groups can be created by spiralling two test forms or randomly assigning examinees to forms. In this design, each group takes only one of the two test forms; thus eliminating any potential practice effect. In addition, because the groups are randomly equivalent, no group effect should be present.. This design is also more economical than the single-group design. Data on both test forms can be gathered in the amount of time it takes to administer a single test form.

Creating the two test forms. The type of test that each group takes can be more complicated when using bilinguals than when performing a two-group equating design. The most straightforward option is to have one group take the "language A" form, and the other group take the "language B" form. Although this option parallels the two-group equating situation (each group takes an intact test or anchor form), it is not optimal when testing bilinguals. Using this design, the performance of the first group in language B cannot be evaluated, nor can the performance of the second group in language A. A better alternative may be to have each group take a hybrid form that contains items in both language A and language B.

An example of this type of mixed-language administration design is the study conducted by Berberoğlu and Sireci (1996). This study evaluated the translation fidelity of two sets of items from two versions of a teacher evaluation form. The original version of this test was in Turkish and the adapted version was in English. To control for the practice effect, examinees answered the items in only one language. However, both Turkish and English items appeared on each of the two test forms. This was accomplished by alternating between the two languages on each form. On the first form, all odd-numbered items were in English and all even-numbered items were in Turkish. The reverse pattern occurred on the second form. For example, the Turkish version of item number one on the first form was item number one on the second form; the English version of item number two on the second form was item number two on the first form, etc. In addition, two English items were included on each form. These items provided an anchor that was used within an IRT analysis to verify that the assumption of randomly equivalent groups was appropriate.

The design used by Berberoğlu and Sireci gathered data in both groups on both languages. However, the effect of alternating between languages on examinees' performance is unknown. An alternative strategy is to have separate sections of the test for each language. Unfortunately, this strategy may introduce an item-order effect in addition to language effect. Interviewing the bilingual examinees may help determine whether alternating the language ordering of the items was confusing or impeded their performance in some way.

As noted above, a serious limitation of the Berberoğlu and Sireci study was that only one type of bilingual group was used. The bilingual sample comprised students at a Turkish university where English was the primary language of instruction. Although they did screen out students who self-reported themselves as "poor" in reading or understanding English, this design did not include any bilinguals whose first language was English. A more thorough evaluation of translation fidelity would include both English-Turkish as well as Turkish-English bilinguals. Thus, the two-group bilingual design can be improved by including more than one type of bilingual examinee.

The Four Group Design

An obvious addition to the two-group bilingual design is two have two groups of bilinguals, who differ with respect to native language, take each test form. The first group would comprise bilinguals who are dominant in the first language and the second group would comprise bilinguals who are dominant in the second language. Individuals in each group would be assigned to one of the two (preferably mixed-language) test forms. In addition to ensuring a more representative group of bilinguals, this design allows for the analysis of performance differences between the two types of bilinguals. DIF and dimensionality analyses could be conducted separately for each group. For example, if an item appears statistically equivalent for both Turkish-English and English-Turkish bilinguals, further evidence is gathered that the item is "the same" in both languages. If an item exhibits DIF in one bilingual group, but not in the other bilingual group, information is gathered pertaining to the different linguistic interpretations of the item. This extended design provides increased information regarding the interaction between the native language orientation of bilinguals and language of the item.

Strategies for Improving Translation DIF Studies

The three issues discussed above illustrate primary factors that must be considered when evaluating dual language exams. First, the confounding between language group effects and test language effects must be addressed. Second, the dimensionality of each language version of the test, and the comparability of the dimensions across languages must be determined. Finally, appropriate data collection and analysis designs must be used. Given the problems inherent in evaluating dual language tests, it appears that the design should incorporate groups of bilingual examinees whenever possible, and use multidimensional approaches to DIF evaluation.

One limitation of previous research for evaluating translation DIF is that unidimensional DIF detection models were used. As noted above, it is likely that different language versions of a test may be measuring more than one dimension.

The comparability of the construct measured across languages could be investigated using multi-group factor analyses; but isolation of a common factor, which would serve as the conditioning variable for subsequent translation DIF analyses, is probably not possible using factor analytic procedures. A more sensible approach is to use a multidimensional DIF detection model. For example, a multidimensional approach to translation DIF could be used to adjust for language proficiency effects that may be confounded with proficiency measured by the test (Sireci, in press).

Multidimensional IRT models for DIF detection (e.g., Ackerman, 1994) are theoretically appealing in this situation, but unfortunately, practical procedures for using these models to evaluate DIF are currently impracticable (Swaminathan, 1994). However, logistic regression DIF detection procedures are particularly applicable to the problem of detecting DIF within a multidimensional context (Clauser, Nungester, Mazor, & Ripkey, 1996; Mazor, Kanjee, & Clauser, 1995; Swaminathan & Rogers, 1990).

For example, a research design that includes two groups of bilinguals who differ with respect to their native language, and includes assessments of proficiency in both languages, would provide the data for a multidimensional logistic regression analysis (assuming the test items are dichotomously scored). In this manner, three proficiency estimates can be entered into the logistic regression equation for each examinee: a proficiency estimate for the characteristic measured by the test, an estimate of the examinee's native language proficiency, and an estimate of the examinee's second language proficiency. Thus, this three-dimensional model would represent a multivariate matching criterion more suitable to the translation DIF situation. The studied items would represent one dimension and the measures of language proficiency would represent two other dimensions to be used as covariates in evaluating the translation DIF of the studied item.

Obviously, this is an immodest proposal. Obtaining adequate samples of two types of bilinguals for a DIF study is a formidable task. Therefore, studies using simulated data should be conducted first. However, generating these multivariate criteria and specifying differences between all relevant proficiency distributions is an arduous task (perhaps there is a dissertation study here!). For example, the effects of differences between the proficiency distributions on all three factors, and the interaction of such effects, needs to be determined under several conditions. The effects of sample size differences between the groups and items is another important issue to be studied. The results of studies using simulated data will help gauge the efficacy of the logistic regression procedure for evaluating translation DIF, if such data can be collected. In any event, extending the traditional DIF research design to include at least two groups of bilingual examinees, perhaps in addition to two groups of monolingual cohorts, and using the logistic regression DIF detection procedure, appears to be an attractive method for evaluating translation DIF.

Conclusion

There is a "big Dif" between DIF evaluations in one language and analysis of translation DIF. Evaluating the statistical equivalence of items translated from one language to another is a far more complex task than evaluations of DIF across groups operating within a single language. Current DIF applications in this area are extremely limited. However, by extending the data collection design to include multiple groups and measures of language proficiency, more promising models for evaluating translation DIF are possible.

References

Ackerman, T. A. (1994). A discussion of measurement direction in a multidimensional latent space and the role it plays in bias detection. In D. Leveault, B. Zumbo, M.E. Gessaroli, & M. Bosss (Eds.). Modern theories of measurement: problems and issues, (pp.105-140), Ottowa, Canada, Edumetrics Research group, Univeristy of Ottowa.

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.

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.

Berberoğlu, G. , & Sireci, S.G., & Hambleton. R.K. (1997). Comparing translated items using bilingual and monolingual examinees. Laboratory of Psychometric and Evaluative Research Report No. Amherst, MA: University of Massachusetts, School of Education.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. Applied Psychological Measurement, 12, 261-280.

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.

Budgell, G. R., Raju, N .S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19, 309-321.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. Journal of Educational Measurement, 33, 202-214.

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

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

Fraser, C. (1983). Noharm II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, University of New England, Centre for Behavioral Studies.

Geisinger, K.F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6, 304-312.

Green, B.F. (1994). Differential item functioning: Techniques, findings and prospects. In D. Leveault, B. Zumbo, M.E. Gessaroli, & M. Bosss (Eds.). (pp. 141-162). Modern theories of measurement: problems and issues, Ottowa, Canada, Edumetrics Research group, Univeristy of Ottowa.

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., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological Assessment, 11, 147-157.

Holland, P.W., & Wainer, H. (Eds.), Differential item functioning. Hillsdale, New Jersey: Lawrence Erlbaum.

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.

Mazor, K. M., Kanjee, A., & Clauser, B. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. Journal of Educational Measurement, 32, 131-144.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrica Monograph Supplement, 4, Part 2, Whole No.17.

Sireci, S.G. (1997). Problems and issues in linking assessments across languages. Educational Measurement: Issues and Practice, 16 12-19.

Sireci, S.G. (1996, August). Using bilinguals to evaluate the comparability of different language versions of a test. Paper to be presented at the annual meeting of American Psychological Association meeting in Toronto.

Swaminathan, H. (1994). Differential item functioning: A discussion. In D. Leveault, B. Zumbo, M.E. Gessaroli, & M. Bosss (Eds.). Modern theories of measurement: problems and issues, (pp. 171-180), Ottowa, Canada, Edumetrics Research group, Univeristy of Ottowa.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Thissen, D, Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland, & H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum.

Valdés, G., & Figueroa, R.A. (1994). Bilingualism and testing: A special case of bias. Norwood, NJ: Ablex.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029590

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Evaluating Translation Equivalence: So What's the Big Dif?

Author(s): Sireci, SG; Swaminathan, H

Corporate Source: University of Mass. Amherst

Publication Date: 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 media, and sold through the ERIC Document Reproduction Service (EDRS). 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 and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic 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.

Sign here,→ please

Signature:

Printed Name/Position/Title: Stephen G. Sireci

Organization/Address:
University of Massachusetts
School of Education
Hills South
Amherst, MA 01003

Telephone: 413 5450564

FAX:

E-Mail Address: sireci@acad.umass.edu

Date: 2/5/97

(over)

ERIC® **Clearinghouse on Assessment and Evaluation**

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

September 11, 1998

Dear ERIC Contributor:

The ERIC Clearinghouse on Assessment and Evaluation is pleased to send you the enclosed resume and microfiche of your publication.

The resume appears in the April 1998 issue of Resources in Education (RIE), the monthly abstract publication of the Educational Resources Information Center (ERIC).

Copies of your document will be available in microfiche in more than 800 institutions around the world. If appropriate, your document may also be purchased in microfiche and paper copy by calling or mailing the enclosed EDRS order form to:

ERIC Document Reproduction Service (EDRS)
7420 Fullerton Road, Suite 110
Springfield, VA 22153-2852
(800) 443-ERIC (3742) or (703) 440-1400

Thank you for your contribution. For your convenience, I have enclosed a copy of the ***"new"*** Reproduction Release form for future submissions. Please submit the Release form with submission(s) to this letterhead address.

Sincerely,

Laura Chapman

Laura Chapman
Sr. Production Specialist

Enclosure:
- Document Resumes(s)
- Microfiche
- EDRS Order Form
- Reproduction Release Form"new"