DOCUMENT RESUME

ED 430 995 TM 029 821

AUTHOR Price, Larry R.
TITLE Differential Functioning of Items and Tests versus the Mantel-Haenszel Technique for Detecting Differential Item Functioning in a Translated Test.
PUB DATE 1999-04-00
NOTE 29p.; Paper presented at the Annual Meeting of the American Alliance of Health, Physical Education, Recreation, and Dance (Boston, MA, April 12-16, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Certification; *Item Bias; Licensing Examinations (Professions); Tables (Data); Test Construction; *Test Items; *Translation
IDENTIFIERS Item Bias Detection; *Mantel Haenszel Procedure

ABSTRACT

Data from a 50-item translated test used for certification were used to assess the percentage and type of agreement between the Mantel-Haenszel (MH) and Differential Functioning of Items and Tests (DFIT) techniques for the detection of differential item functioning (DIF). The DFIT procedure flagged 10 of 30 items as exhibiting significant DIF while the MH technique flagged 2 of 30 items for significant DIF. In both methods items were flagged for significant DIF when translation differences appeared in the item stems. The DFIT method was more sensitive in detecting DIF, resulting exclusively from differences in the item answer options. The overall percent agreement between the two techniques for the detection of DIF in this investigation was 20 percent. The MH technique detected 1 of 10 items as exhibiting nonuniform DIF and 1 of 10 items as displaying uniform DIF. The DFIT procedure detected 4 of 10 items as exhibiting nonuniform DIF and 6 of 10 as displaying uniform DIF. Four appendixes contain tables of descriptive statistics and the English and back-translated item versions. (Contains 34 references.) (Author/SLD)

Running Head: DFIT VS. MH FOR DETECTING DIF IN TRANSLATIONS

Differential Functioning of Items and Tests Versus the Mantel-Haenszel Technique for Detecting Differential Item Functioning in a Translated Test

Larry R. Price

Emory University

Paper Presented at the Annual Meeting of the American Alliance of Health, Physical Education, Recreation and Dance, Boston, MA.

Abstract

Data from a 50-item translated test used for certification were used to assess the percentage and type of agreement between the Mantel-Haenszel (MH) and Differential Functioning of Items and Tests (DFIT) techniques for the detection of differential item functioning (DIF). The DFIT procedure flagged 10/30 items as exhibiting significant DIF while the MH technique flagged 2/30 items for significant DIF. In both methods items were flagged for significant DIF when translation differences appeared in the item stems. The DFIT method was more sensitive in detecting DIF, resulting exclusively from differences in the item answer options. The overall percent agreement between the two techniques for the detection of DIF in this investigation was 20%. The MH technique detected 1/10 items as exhibiting non-uniform DIF and 1/10 items displaying uniform DIF. The DFIT procedure detected 4/10 items as exhibiting non-uniform DIF and 6/10 as displaying uniform DIF.

Introduction

Translated assessment instruments used for the purpose of certification and licensing are sometimes translated from one language to another when they are used in a cross-cultural setting. Examples of such instruments in physical education and exercise science include certification tests administered to practitioners for the purpose of knowledge mastery in a discipline prior to becoming certified to perform the clinical assessment of physical fitness/wellness or for participation in certain types of sports activities. When tests are modified and used cross-culturally, the measurement equivalence of the instrument should be evaluated. If measurement inequivalence is found, the test should be revised by improving or replacing problematic items.

The original and modified tests may not be equivalent because: (a) through the translation process the meaning of the test items have been unknowingly changed and/or (b) the test items may not have the same relevance across the different cultural groups (Budgell, Raju, & Quartetti, 1995). Historically, cross-cultural researchers have used procedures, such as back-translation and decentering, as an initial step in the process of test translation (Brislin, 1980). After test translation was complete, classical test theory methods were used for examining differences within groups with the final goal of producing measurement equivalence across groups.

Classical test theory methods are population or group dependent, however, and are therefore less than ideal for verifying measurement equivalence in translated tests. Statistical methods based on item response theory (IRT) overcome a variety of problems associated with the classical test theory model and provide researchers with an improved methodology for examining measurement equivalence across culturally different groups.

Within the framework of IRT, measurement equivalence is a property that exists when the relations between observed test scores and the latent attribute measured by the test are identical across sub-populations (Drasgow, 1984, p. 134). In order for a translated test to exhibit measurement equivalence, individuals who come from different cultural groups that are equal in ability must have the same observed score. Equivalent assessment instruments must be used in

cross-cultural testing in order to determine if true score differences exist across groups or if score differences are attributable to measurement inaccuracies.

A summary of related literature reveals a limited number of studies that have been conducted using IRT-based DIF analyses to assess the measurement equivalence of translated tests (Hulin, Drasgow, & Komocar, 1982; Drasgow & Lissak, 1982; Candell & Roznoswki, 1984; de Vera, 1985; Osberg, Scott, & Raju, 1985; Drasgow & Hulin, 1984; Hulin & Mayer, 1986; Candell & Hulin, 1987; Ellis, 1989; Ellis, Minsel, & Becker, 1989; Ellis, 1991; Budgell, Raju, & Quartetti, 1995; Price & Oshima, 1998). Those studies have not provided consistent and conclusive evidence regarding the effectiveness of IRT-based DIF analysis to assess the measurement equivalence of translated tests. The dissimilar results among those studies may be due to methodological restrictions, such as the confounding effect of the linguistic ability of the two groups, the quality of the test translation process, or a small sample size.

The first purpose of this study was to conduct additional empirical work on the application of Differential Functioning of Items and Tests (DFIT) for the detection of DIF in a translated test as proposed by Raju, van der Linden, and Fleer (1992). The IRT-based DFIT procedure was selected for use in this study because it offered several appealing measures including an overall measure of differential test functioning (DTF), compensatory differential item functioning (C-DIF), and non-compensatory differential item functioning (NC-DIF).

A second purpose of this study was to assess and compare the agreement of the Mantel-Haenszel (MH) technique (Holland & Thayer, 1988) with the DFIT procedure for the detection of differential item functioning (DIF). Additionally, the effectiveness of the Mantel-Haenszel (MH) technique for assessing the measurement equivalence of translated tests has little previous empirical evaluation. Using the MH technique to model and identify DIF in translated tests avoids many of the practical problems related to IRT-based methods, such as obtaining a large sample size for minority groups and using special software for DIF analysis.

The Mantel-Haenszel Technique and the Detection of Differential Item Functioning

Holland and Thayer (1988) proposed that an observed-score-based statistical technique developed by Mantel and Haenszel (1959) be used for the detection of DIF in educational settings. The application of the MH technique for the detection of DIF involves the formation and comparison of two groups - a reference group against which a focal group is compared. Most often, the focal group is the culture or gender group that is of interest in terms of detecting DIF. Prior to analysis, the two comparison groups must be matched on some criterion variable directly related to the construct being measured. The criterion most often used for matching the ability of the subjects in the focal and reference groups is the total test score. Once the two groups are matched on ability, the performances of the two groups are analyzed item by item. An odds ratio, α, is computed for the ith item through

$$\alpha_i = \frac{p_{ri} q_{fi}}{p_{fi} q_{ri}}, \qquad (1)$$

where p_{ri} is the proportion of subjects in the reference group that responded correctly to the ith item, q_{ri} is the proportion of the reference group that responded incorrectly, p_{fi} is the proportion in the focal group that responded correctly, and q_{fi} is the proportion in the focal group that responded incorrectly. The odds ratio, α, is an estimate of DIF effect size, and the values of α_i, range from 0 to ∞ with an α_i of 1 indicating null DIF. The odds ratio metric is not particularly meaningful to test developers who are used to working with numbers displayed in an item difficulty scale. Therefore, the odds ratio may be converted to log odds because the latter is symmetric around zero and easier to interpret. An α_i value greater than 1 may indicate that the item functions differentially across the two groups in favor of the reference group. The MH - χ^2 is distributed approximately as a chi-square statistic with 1 degree of freedom.

Advantages of using the MH technique for the detection of DIF include its computational simplicity and associated test of statistical significance (Bugdell, Raju, & Quartetti, 1995; Rogers

& Swaminathan, 1993; Raju, Drasgow, & Slinde, 1993). Additionally, the technique does not require a large sample size, as is the case when implementing IRT-based DIF detection methods. One disadvantage to using the MH technique is its insensitivity in detecting non-uniform DIF across ability levels. Non-uniform DIF occurs when the differences in probabilities for success on an item are not uniform across ability levels for the two groups under study.

Item Response Theory and the Detection of Differential Item Functioning

Although the detection of DIF can be accomplished using a variety of techniques, Angoff (1993) suggested that the methods of choice are the MH procedure, which permits the examination of item difficulty effects, and the comprehensive three-parameter IRT approach, which permits the observation of differences in any of the characteristics of the IRT function. Recently, Bugdell, Raju, & Quartetti (1995) conducted a study using the MH, Lord's Chi-square Test, and Signed and Unsigned Area methods for the detection of DIF on a translated test used for certification and licensing. In that study, the percentage of agreement between Lord's Chi-square and the MH technique for the detection of DIF was substantial, while the MH technique consistently identified more items with significant DIF than any of the three IRT-based methods.

In past studies, judgmental and empirical evaluations of DIF items were conducted in order to explain the possible cause of DIF. After DIF items were evaluated by a bilingual test content committee, test items were corrected to eliminate DIF and either returned to the item pool or eliminated from the test entirely. Theoretically, the end result was a test that exhibited measurement equivalence.

In IRT, the probability of a correct response on an item for an individual with a latent trait θ is described by an item characteristic function (ICF), the S-shaped curve. Additionally, the curve is often, but not always, defined by three parameters and represented by the logistic function

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i}(\theta - b_i)}{1 + e^{Da_i}(\theta - b_i)} \qquad (2)$$

where $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly, b_i is the item difficulty parameter, a_i is the item discrimination parameter, c_i is the pseudo-chance-level parameter, and D is a scaling factor designed to make the logistic function closely approximate the normal ogive function (D = 1.702).

The cornerstone of IRT is based on the property of invariance. This property implies that the ICF is unique under the conditions of a particular model except for random variations after item parameters are placed on a common scale. Linn et al. (1981) and Shepard (1987) agreed that when the curves from the two groups were different, the fundamental assumptions of item response theory models were violated. The assumptions of the item response theory model include unidimensionality (e.g., that the test measures only one construct) and local independence. Evidence of large DIF indicates that an item is measuring an additional construct (possible multidimensionality) in one of the two groups, and the construct may not be relevant to the intended purpose of the test.

Differential Functioning of Items and Tests (DFIT)

Raju et al. (1992) proposed a parametric framework, known as Differential Item Functioning of Items and Tests (DFIT), that allowed for individual DIF to add up to total test differential test functioning (DTF). Because the test item is the most fundamental part of a test, DIF studies at the item level are important in uncovering possible unfairness in test use. However, it is possible for several items within a test to exhibit DIF without the overall test being unfair. Therefore, potential unfairness at the test level (DTF) should also be examined. The following paragraphs briefly describe the DFIT framework (Raju et al., 1995).

The Differential Functioning of Items and Tests (DFIT) Framework

Differential test functioning. Differential Functioning of Items and Tests begins with a measure of differential test functioning (DTF). Within DFIT, $P_i(\theta s)$ represents the probability of

success for an examinee s with ability θ on item *i* (Equation 1). The test may consist of *k* items and have one set of item parameters for each of two groups (reference group and focal group). Further, an assumption is made that the two sets of item parameters are on a common scale. The probability of success, $P_{iR}(\theta s)$, on item *i* for examinee s is expressed as if he or she were a member of the reference group. Similarly, $P_{iF}(\theta s)$ represents the same probability of success for the same examinee on the same item as if he or she were a member of the focal group. If an item is functioning differently in two groups, P_{iR} and P_{iF} should be different for some examinees.

Within IRT, an examinee's true score can be expressed as

$$T_s = \sum_{i=1}^{k} P_i(\theta_s). \qquad (3)$$

Theoretically, in this explanation, each examinee will have two true scores, one as a member of the reference group (T_{sR}) and the other as a member of the focal group (T_{sF}). If T_{sR} and T_{sF} are equal for an examinee, the examinee's true score is independent of membership. Furthermore, the greater the difference between T_{sR} and T_{sF}, the greater the differential functioning of a test. A measure of DTF at the examinee level may be defined as $(T_{sF} - T_{sR})^2$. Therefore, an overall measure of DTF across examinees may be defined as

$$\text{DTF} = \underset{F}{\in} (T_{sF} - T_{sR})^2, \qquad (4)$$

where the expectation (∈) is in reference to the focal group. Next, allowing that $Ds = T_{sF} - T_{sR}$, Equation 4 can be rewritten as

$$\text{DTF} = \in D_s^2 = \int_\theta D_s^2 f_F(\theta) d\theta = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 \sigma_D^2 + \mu_D^2, \quad (5)$$

Where $f_F(\theta)$ is the density function of θ in the focal group, and μ_{TF} and μ_{TR} represent the expected proportion correct for examinees in the focal and reference groups.

Differential Item Functioning

The DFIT model provides for the calculation of two measures, compensatory DIF (C-DIF) and non-compensatory DIF (NC-DIF). The two measures provide unique but related types of information about the functioning of an item. The formulation of C-DIF includes the covariance between the differences in item probabilities and the difference between the two expected proportions correct for each item on the test. Additionally, the C-DIF measure includes the mean difference in item probability and the mean difference between the two expected proportions correct for each group on each respective item. C-DIF is related to the DTF as follows:

$$\text{DTF} = \sum_{i=1}^{k} \text{C-DIF}_i . \quad (6)$$

Since DTF is the sum of C-DIF, there is a possibility for cancellation of differential functioning at the test level when one item displays C-DIF in favor of one group and another item displays C-DIF for the other group. In practical settings, a test developer can examine the compensating magnitude of items displaying C-DIF, and the construction of the item stem and

distracters after translation, in order to make a decision regarding item removal or revision. Ultimately, items displaying C-DIF can be removed from the test in order to reduce overall DTF.

NC-DIF, on the other hand, assumes all items other than the one under study are free from differential functioning. Therefore NC-DIF is not additive. Two features of NC-DIF are presented. First, since the differences in item probabilities for each item on the test are included in the DFIT model, NC-DIF = 0 if and only if the item parameters for each item are equal for the reference and focal groups. Further, Lord's chi-square test offers a test of the null hypothesis that NC-DIF = 0. In this sense, NC-DIF is similar to other IRT-based DIF indices, such as chi-square (Lord, 1980) and area measures (Raju, 1988). Raju (1992) noted that items having significant NC-DIF do not necessarily have significant C-DIF. An example of this occurs when one item favors the reference group and another item favors the focal group. In this case, NC-DIF occurs, but C-DIF may not be significant. Second, allowing $f_F(\theta)$ to denote the density function of θ in the focal group, NC-DIF may be written as

$$\text{NC-DIF}_i = \int_{-\infty}^{\infty} [P_{iF}(\theta) - P_{iR}(\theta)]^2 f_F(\theta) d\theta. \quad (7)$$

Including the density function into equation 7 not only provides for DIF detection, but also provides a measure of impact or effect size. Equation 7 is a measure of impact identical to a definition offered by Wainer (1993).

DFIT Significance Test

Raju et al. (1995) proposed a significance test for DTF. A significant chi-square for DTF indicates that there exists a significant differential test functioning. When DTF is significant, one item (typically an item with large C-DIF) is removed and DTF is tested again. This process is repeated until the chi-square test shows no significance. Those items removed to achieve non-

significant DTF are regarded as "significant" C-DIF items. Although a significance test for NC-DIF was theoretically described in Raju et al. (1995), the authors recommended an empirical approach to declare the significance of NC-DIF (NC-DIF >.006) based on a simulation study by Fleer (1993).

Method

Item and Test Translation

One bilingual Japanese scuba diving content expert independently translated the 50-item test from English to Japanese and then back-translated the test to English. Although, most of the items in the Japanese version were translated precisely or as closely as possible from items on the English version, some of the items could not be translated word-for-word due to differences between Japanese and English languages. Therefore, the translators identified 30 of the 50 items semantically and linguistically similar enough to be included in this study. The 30-item test included the following six content areas: (a) skills/safety, 13 items; (b) decompression, 1 item; (c) physics, 5 items; (d) physiology, 6 items; (e) equipment, 2 items; and (f) environment, 3 items. Of the 30 items, 17 precisely translated. Of the remaining 13 items, some items had stem and/or option differences but still tapped into the same content area. Separate keys were used for the English and Japanese versions.

Samples

Data collection for this study was conducted during May through October 1996, in Tokyo, Japan and, in the United States, California, Georgia, and Florida. The subjects participating in the study were Japanese and American males and females between the ages of 18 and 40 years enrolled in an entry-level scuba diving certification test sanctioned by the National Association of Underwater Instructors (NAUI). The sample consisted of 1,134 Japanese and 1,000 American males and females. All subjects participating in the study had a minimum of 12 years of formal education in their respective country's educational system. No students with mental or physical disabilities participated in the study. This information was offered on the student's individual confidential course file that was completed prior to the course beginning.

Course Curriculum and Test Administration

The assessment instrument used in this study was a criterion-referenced mastery test used for certification in sport scuba diving. Both groups received a course curriculum written by NAUI. The standard NAUI test for certification was administered to both groups at the end of the formal course of instruction. Subjects were given 1 hour and 30 minutes to complete the 50-item test. Notes or textbooks were not used as reference material during the test, however, calculators were allowed in order to compute applied problems related to diving physics. Students used decompression tables to complete the applied decompression problems on the test. All answers were recorded on a separate answer sheet with an identification number.

Sample Comparisons

Ideally, if IRT- and MH-based procedures are effective in detecting DIF in translated tests, the same items should be identified as having significant DIF by each technique. For each method the total number of items displaying DIF was determined. The number of common DIF items across methods was also determined and expressed as a percentage of agreement between the two methods.

Data Analysis and Parameter Estimation

After data collection of the 50-item test, test answer sheets were examined for errors and accuracy of answer coding. Then, the data were reduced to 30 linguistically and semantically similar items for the DFIT analysis. Finally, parameter estimation and data analysis were conducted in the following sequential steps:

1. The Statistical Package for the Social Sciences personal computer software program (SPSS 7.5) was used to calculate descriptive statistics related to the demographics of the sample and to perform classical item analysis. In addition, the reliability of the test was investigated.

2. Unidimensionality of the test was verified ($\chi^2 = .652$, $p = .326$) by using the DIMTEST (Stout, 1991) computer program.

3. A BILOG 3.10-PC computer was used for the estimation of item and ability parameters. The program's Marginal Maximum Likelihood Estimation (MMLE) procedure was

used for the estimation of item parameters under the three-parameter logistic model. Estimates of underlying ability were made via the program's Bayesian EAP procedure using the unit normal prior. Goodness-of-fit indices generated by BILOG were examined for model-data fit.

4. After the item parameters were estimated, the Japanese and American examinees were placed on a common scale by the test characteristic curve method (Stocking & Lord, 1983) as incorporated into the computer program IPLINK (Lee & Oshima, 1996).

5. DIF and DFIT measures were computed using the framework proposed by Raju et al. (1992).

6. DIF and DFIT indices were computed for theta ranges across the entire ability range. DIF and DFIT indices were computed using the estimated a-, b-, and c-parameters. DIF measures computed included the chi-square statistic for DTF. For all measures, items were examined for significant differential functioning at the alpha level of .01. NC-DIF items were declared significant if they had a value greater than .006. The .006 significance level was empirically established through a previous Monte Carlo study by Fleer (1993).

7. Mantel-Haenszel chi-square statistics were computed for each item using the computer program StatXact 3 (Gajjar, Mehta, Patel, & Senchaudhuri, 1997). A chi-square distribution with 1 degree of freedom and an alpha level of .01 was used for identifying items with significant DIF.

Results

The results of this study are organized into five sections. The first section reports the demographic characteristics and descriptive statistics for the sample. The second section describes the differential test functioning and compensatory DIF results. Section three reports the non-compensatory and MH DIF results. The fourth section reports a comparison of the uniform and non-uniform DIF observed in the analysis. The final section provides the results of a review conducted by a committee of expert scuba diving instructors.

Demographic Characteristics of the Sample

The demographic characteristics included 1,000 American and 1,134 Japanese males and females between the ages of 18 and 40 years. Appendix A provides a descriptive summary for the

sample used in this study. Appendix B provides classical item statistics for the Japanese and American samples.

Differential Test Functioning and Compensatory DIF Results by Item Content

DTF was significant (χ^2 = 7015.45, p<.0001), indicating that the two versions of the 30-item test were functioning differentially at the test level. Item number 30 from the skills/rescue content area was found to have the greatest amount of C-DIF (.160). After elimination of item 30, the chi-square statistic for DTF was not significant (χ^2 = 1205.97, p > .05. Therefore, the only significant item C-DIF item was item 30. Appendix C provides the selected output from the DFIT program along with results from the MH analysis.

Non-Compensatory DIF (NC-DIF) and Mantel-Haenszel DIF Results

The following items (five skills/safety, three physiology, and two physics) were found to have significant NC-DIF values (NC-DIF >.006): 3, 4, 11, 12, 13, 16, 22, 25, 29, and 30. NC-DIF assumes that all other items in the test are free of DIF and therefore does not include information about DIF from other items. Therefore, NC-DIF values are particularly good for revealing why certain items exhibit more DIF than others or why various items may be offensive to certain groups. Mantel-Haenszel statistics may be compared directly with NC-DIF values since the MH assumes no relationship with other items on the test. The MH technique flagged items 4 and 30 as having significant DIF. Thirteen out of 30 of the items displayed translation differences either in the stem and/or in the options. Six of those 13 items were identified as having significant NC-DIF while only 2 of the 13 items displayed significant DIF, according to the MH technique. That amount of DIF identification translated into an agreement between the two procedures of 20%. Appendix D includes the text for items 4, 13, and 30 in their original English form as well as the text in the Japanese version after translating from Japanese to English word-by-word. Items 4, 13, and 30 were selected for illustration of DIF detection using the DFIT and MH procedures because those items exhibited the greatest NC-DIF, and item 30 showed large C-DIF as well.

Uniform and Non-Uniform DIF

Uniform DIF occurs when the difference in probabilities of success is uniform for the two groups overall ability levels. In this study, items 3, 4, 12, 22, 25, and 29 exhibited uniform DIF.

Non-uniform DIF occurs when the probability of success is greater for one group at one end of the ability scale and the probability of success is greater for the other group at the other end of the ability scale. The item characteristic curves for the two groups cross at some point when graphically examined. Items that exhibited non-uniform DIF in this study were 11, 13, 16, and 30. Appendix E identifies how items are classified in relation to both types of DIF.

Review by Scuba Diving Experts

An additional way to provide further analysis related to comparison between the two statistical methods is to consider information provided by a translation review committee. Three bilingual content experts in sport scuba diving independently evaluated the translation quality according to a modified translation evaluation scale adapted from Budgell, Raju, and Quartetti (1995). Appendix F provides a summary of the results from the rating exercise. Nine out of 10 of items that were identified as having significant DIF by a statistical procedure were also identified by the content/translation raters as being problematic. The items (4 and 30) displaying the greatest DIF had an average rating score by the evaluators of 2 and 3, respectively. In this study, the results from the content/translation evaluation exercise provided additional support for the effectiveness of both statistical methods for the detection of problematic items.

Discussion

This study provides a comparison of the DFIT framework and the MH technique for the detection of DIF on a test administered for certification. Specifically, this study provides evidence that measurement equivalence is often difficult to achieve in translated tests used in cross-cultural settings. The DFIT procedure flagged 10 out of 30 items for significant NC-DIF while the MH procedure flagged only two items for significant DIF. Both of the methods had an agreement of 20%. This pattern of DIF identification is in agreement with some studies, but is in

disagreement with others. This lack of consensus between related studies might be due to the following limitations of this study.

First, the items on both tests displayed a low level of discrimination. Although this study provides evidence that the MH technique is effective for identifying significant DIF on test on poorly discriminating test items, these items confound the process of deciphering the possible true cause of DIF. Second, the translation process was not performed with a high degree of accuracy. This translation accuracy problem was evident upon review of the results in Appendix F. Both the DFIT and MH procedures flagged items that had stem translation differences, and the DFIT procedure was more sensitive than the MH technique at identifying both uniform and non-uniform DIF.

Interestingly, the MH technique detected one case of non-uniform DIF in this study. Typically, the MH technique is not sensitive to this type of DIF. One possible explanation regarding why the MH technique flagged item 30, a non-uniform DIF item, could have been because the item characteristic curves for the two groups functioned in a uniformly differential manner across practically the entire ability range.

For test developers involved in cross-cultural testing situations, the non-IRT-based MH technique is easy to implement using standard statistical packages and may be used with small sample sizes. These are important practical considerations in applied test development and evaluation settings. However, even though using the DFIT framework to detect DIF may be more difficult logistically, the results of this project indicate that the increased difficulty may be well worth the effort when the detection of subtle differences in culture is an important consideration. Additionally, the DFIT framework allows for a measure of DTF regarding the overall functioning of test fairness while the MH technique does not. Finally, DIF detection is an important first step in an analytical process of determining the cause and explanation for such DIF. Expert evaluation of item and test DIF/DTF obtained through independent translator ratings is an important follow-up step in order to merge important information from statistical methods with expert content evaluation.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 1-29). New Jersey: Lawrence Erlbaum Publishers.

Brislin, R. (1980). Translation and content analysis of oral and written material. In H. C. Triandis and J. W. Berry (Eds.), Handbook of Cross-Cultural Psychology (Vo. 2, pp. 389-444). Boston: Allyn and Bacon.

Bugdell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19 (4) 309-321.

Candell, G. L., & Hulin, C. L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology, 17, pp. 417-440.

Candell, G. L., & Rozonowski, M. (1984). Using IRT to Establish Equivalence Across U.S. and Canadian Sub-populations. Paper presented at the annual meeting of the American Psychological Association, Toronto.

de Vera, M.V. (1985). Establishing Cultural Relevance and Measurement Equivalence Using Emic and Etic Items. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. Psychological Bulletin, 95, 134-135.

Drasgow, F., & Lissak, R. I. (1982). Modified paralell analysis: A procedure for examining the latent dimensionality of dichotomously-scored item responses. Cited in C. L. Hulin, F. Drasgow, & C. K. Parsons. Item Response Theory: Application to Psychological Measurement. Homewood, IL: Dow Jones-Irwin.

Ellis, B. (1991). Item Response theory: A tool for assessing the equivalence of translated tests. International Test Bulletin, 32-33, 33-51.

Ellis, B. B. (1989). Differential item functioning: implications for test translations. Journal of Applied Psychology, 74 (6), 912-921.

Ellis, B.B., Minseal, B. & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. International Journal of Psychology, 24, 661-684.

Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of items and test bias. (Doctoral dissertation, Illinois Institute of Technology). Dissertation Abstracts International, 54-04, 2266B.

Gajjar, Y., Mehta, C., & Patel, N. (1997). StatXact (computer program). Cambridge, MA: CYTEL Corporation.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations. Journal of Applied Psychology, 67.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18. pp. 115-142.

Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the job description index into Hebrew. Journal of Applied Psychology, 71. pp. 83-94.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lee, K., & Oshima, T. C. (1996). IPLINK. (computer program). Atlanta: Georgia State University.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5. 159-173.

Lord, F. (1980). Applications of Item Response Theory to Practical testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement, (3rd ed., pp. 13-103). New York: MacMillian.

Osberg, D. W., Scott, J. C., & Raju, N. S. (1985). An Anaysis of the Use of Item Response Theory to Investigate the Fidelity of Test Translations. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Price, L., & Oshima, T. C. (1998). Differential Item Functioning and Language Translation: A Cross-National Study with a Test Developed for Certification. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Raju, N. (1988). The area between two item characteristic curves. Psychometrika. 53 (4), 495-502.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992). An IRT-based internal measure of test bias with applications for differential item functioning. Paper presentation at the American Educational Research Association, San Francisco.

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, LordÆs Chi-Square Test and the Mantel-Haenszel technique for assessing differential item functioning. Educational and Psychological Measurement, 53 (2), 301-314.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 19 (4), 353-368.

Rogers, H. J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17 (2), 105-116.

Shepard, L. A. (1987). The case for bias in tests of achievement and scholastic aptitude. In S. Modgil & C. Modgil (Eds.), Arther Jensen: Concensus and controversy (pp. 170-190). New York: Falmer Press.

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

Stout, W. (1991). DIMTEST (computer program). Champaign: University of Illinois.

Wainer, H. (1993). Model-Based Standardized Measurement of an Item's Differential Impact. In P. Holland and H. Wainer (Eds.), Differential Item Fucntioning (pp. 123-135). Hillsdale: Lawrence Erlbaum Associates.

Appendix A

Descriptive Statistics and Reliability for Sample

Group/Gender	n	Mean	SD	Reliability
American	1,000	26.7	1.9	.37
Male	422	26.7	2.0	
Female	578	26.7	1.7	
Japanese	1,134	26.9	1.9	.47
Male	630	27.0	1.9	
Female	504	26.8	2.0	

Note. Coefficient Alpha was used in the computation of the reliability index.

Appendix B

Summary of Classical Item Analysis for Japanese and American Samples

	Japanese		American	
Item	% Correct	Biserial Correlation	% Correct	Biserial Correlation
1	.990	.285	.968	.010
2	.991	.210	.999	-.153
3	.896	.323	.994	.388
4	.893	.060	.722	.183
5	.990	.086	.986	.235
6	.974	.102	.949	.253
7	.987	.001	.969	.186
8	.981	.347	.979	-.025
9	.976	.037	.971	-.056
10	.938	.321	.961	.240
11	.902	.214	.883	.409
12	.811	.139	.729	.077
13	.674	-.055	.557	.202
14	.913	.030	.913	.193
15	.908	.444	.948	.173
16	.833	.318	.928	.173
17	.926	.260	.927	.013
18	.930	.357	.958	.346
19	.932	.147	.903	.142
20	.945	.304	.945	-.139
21	.843	.248	.795	.065

(table continues)

22	.824	.111	.678	.144
23	.906	.174	.827	.115
24	.966	.248	.918	.090
25	.969	.135	.818	.194
26	.935	.324	.907	.019
27	.915	.130	.846	.065
28	.938	.237	.994	.624
29	.887	.416	.974	.000
30	.856	.173	.917	.012

Appendix C

Differential Functioning of Items and Tests and Mantel-Haenszel Statistics

Item	C-DIF	NC-DIF	MH Alpha	Chi-Square
1	-.006	.000	1.0	0.9
2	.002	.000	1.0	0.0
3	.022	.017*	0.2	1.0
4	-.041	.025*	1.4	4.2*
5	.001	.000	0.6	0.3
6	-.001	.001	1.5	0.2
7	-.002	.000	0.4	0.7
8	.000	.000	1.0	0.0
9	.000	.000	0.3	0.1
10	.009	.001	0.5	0.7
11	.015	.008*	1.0	0.0
12	-.008	.008*	1.2	0.5
13	.022	.036*	0.9	0.6
14	.010	.002	0.9	0.2
15	.009	.004	0.6	0.1
16	.037	.012*	0.1	0.0
17	.004	.000	1.6	1.2
18	.010	.001	0.4	1.0
19	-.002	.001	0.8	0.6
20	-.005	.001	0.7	0.1
21	-.008	.001	0.9	0.3
22	-.028	.019*	1.0	0.1
23	-.014	.006*	1.0	0.2

(table continues)

24	-.012	.002	0.4	3.0
25	-.039	.019*	0.7	0.6
26	-.009	.001	1.0	0.0
27	-.016	.003	0.8	0.4
28	.016	.004	0.2	1.2
29	.024	.012	2.4	1.0
30	.160	.271*	2.6	5.3*

Note. 1134 examinees.

Appendix D

English and Back-Translated Japanese Versions

Item	Language	Question Stem
4	English	The least desirable dependent option i an out-of-air situation is buddy breathing. *a) T b) F
4	Japanese	During the emergency procedure for out of air, the most recommended method for getting assistance coming up is to get the optional second slage. *a) T b) F
13	English	To maintain neutral buoyancy during descent, a diver wearing a wet suit should: *a) add air to the Buoyancy Compensator. b) dump all the air from the Buoyancy Compensator. c) activate the J-valve. d) remove some lead from the weight belt.
13	Japanese	During descent, a diver wearing a wet suit should _____to maintain neutral buoyancy. *a) add air to the Buoyancy Comensator. b) dump all the air from the Buoyancy Compensator. c) hold a rock instead of a weight. d) get rid of all of the air in the lungs and hold your breath.
30	English	It is good practice for diving buddies to: a) wear matching equipment. b) have the same certification level. *c) agree on a dive leader. d) practice emergency swimming ascents.
30	Japanese	Which is wrong concerning the buddy system? a) Go down and up together always. b) Swim side by side in the distance that you can reach each other by hand. *c) In the water, one goes first and another dives afterwards. d) Decide beforehand which one will take the leadership.

Note. * Indicates the correct answer choice.

Appendix E

DIF Identification and Classification

Item--Content	DIF	Translation Result	Method
3--skills/safety	uniform	no difference	DFIT
4--skills/safety	uniform	misspelling/negative vs. positive stem	DFIT/MH
11--physiology	non-uniform	no difference	DFIT
12--skills/safety	uniform	different answer choices	DFIT
13--physics	non-uniform	different answer choices	DFIT
16--physiology	non-uniform	different answer choices	DFIT
22--physics	uniform	different answer choices	DFIT
25--skills/safety	uniform	different stem/different answer choices	DFIT
29--physiology	uniform	no difference	DFIT
30--skills/safety	non-uniform	negative vs. positive stem/different answer choices	DFIT/MH

Appendix F

Translation Evaluation Results

	Ratings		
Item	Rater 1	Rater 2	Rater 3
3	4	4	4
4*	3	3	3
11	3	3	3
12	3	3	3
13	3	3	3
16	4	4	3
22	3	3	3
25	4	3	3
29	4	3	3
30*	2	2	2

Note. 1 = poor: meaning of the translation not clear in some places; many stem or distracter differences; may contain some culturally inappropriate material. 2 = fair: good translation in terms of meaning; some stem or distracter differences; contains some linguistic or working errors. 3 = good: equivalent translation in terms of meaning but contains a few linguistic or wording errors. 4 = excellent: equivalent translation in terms of meaning and contains no linguistic errors or culturally inappropriate material.

*Items flagged for significant DIF by DFIT and MH techniques.

TM029821

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

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: "Differential Functioning of Items and Tests Versus the Mantel-Haenszel Technique for Detecting Differential Item Functioning in a Translated Test"

Author(s): Larry R. Price

Corporate Source:

Publication Date: April 25, 1999

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

[X]

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 reproduca 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: L R. P	Printed Name/Position/Title: Larry R. Price, Senior Lecturer
Organization/Address:	Telephone: (404) 727-6527 FAX:
	E-Mail Address: lprice@emory.edu Date: 4/25/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

088 (Rev. 9/97)
PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.