Using bilingual students to link and evaluate different language versions of an exam*

ONG Saw Lan[1], SIRECI Stephen G.[2]

(1. School of Educational Studies, Malaysia Science University, Penang 11800, Malaysia;

2. School of Education, University of Massachusetts, Amherst MA01003, USA)

Abstract: Many researchers and the International Test Commission's (Hambleton, 2005) caution against treating scores from different language versions of a test as equivalent, without conducting empirical research to verify such equivalence. In this study, we evaluated the equivalence of English and Malay versions of a 9th-grade math test administered in Malaysia by conducting several statistical analyses. All analyses were conducted on data from a large sample of English-Malay bilingual students who took both versions of the exam. First, we conducted two equating analyses—one based on classical test theory and another based on item response theory (IRT). Then differential item functioning analyses (DIF) were performed to see if any items functioned differentially across their English and Malay versions. The DIF results flagged 7 items for statistically significant DIF, but only one had a non-negligible effect size. We then conducted another equating analysis dropping the DIF items. The equating results suggested an adjustment of 1 or 2 points, depending on the mathematics achievement levels. The results indicate that bilingual examinees can be useful for evaluating different language versions of a test and adjusting for differences in difficulty across test forms due to translation.

Key words: Differential Item Functioning; linking tests; bilingual examinees

1. Introduction

Developing tests in more than one language is a common approach for educational systems that involve students who operate in different languages. In Malaysia, several national tests are administered in both Malay and English. Scores from these different versions of the exam are treated as equivalent when they are reported and interpreted, but many researchers and the International Test Commission's Guidelines for Adapting Educational and Psychological Tests (Hambleton, 2005) caution against treating scores from different language versions of a test as equivalent, without conducting empirical research to verify such equivalence.

In Malaysia, two language versions of mathematics tests (English and Malay) are administered in the national public examination system to make inferences about grade nine students' mathematic achievement. Achievement level classifications for students are based on raw scores, regardless of which language version of the test was taken. Administering educational tests in different languages requires the comparability of scores

* The finding in this paper has been presented at the International Association for Educational Assessment Annual Conference, September 7-12, 2008, Cambridge, UK.

ONG Saw Lan, Ph.D., School of Educational Studies, Malaysia Science University; research fields: educational measurement, cross-lingual testing.

SIRECI Stephen G., professor, School of Education, University of Massachusetts; research fields: content validity, standard setting, cross-lingual assessment.

between the different language versions of the test evaluated. When differences are found, equating can be used to adjust the scores from different language versions of the test.

The purpose of the present study was to evaluate the equivalence of English and Malay versions of a 9th-grade math test administered in Malaysia. All analyses were conducted on data from a large sample of English-Malay bilingual students who took both versions of the exam. In addition to evaluating the particular test studied, this research evaluated the utility of gathering data from bilingual students for the purposes of evaluating differences in difficulty across test forms due to translation.

2. Test equating and Differential Item Functioning

Test equating is a statistical procedure used in establishing the relationship between scores from two or more tests so that these different tests are placed on a common scale (Kolen & Brennan, 2004). It is often used in situations where examinees taking different forms or different versions of the test are compared to one another. Many researchers believe that a procedure may be called equating only if it is used strictly to equate two test forms that have essentially the same content. In the case of test adaptation or translation, the content is intended to be the same, but since differences in item or test difficulty may occur from the translation/adaptation process, equating procedures may be needed so that scores on the different language versions of the exam can be interpreted interchangeably (Gao, 2004).

Test equating methods can be classified as traditional equating or item response theory equating. Traditional equating methods are based on Classical Test Theory (CTT). In the CTT approach, score correspondence is established by setting characteristics of the score distribution equal for a specified group of examinees (Kolen & Brennan, 2004). Three popular CTT methods are mean equating, linear equating, and equipercentile equating.

Test equating is concerned with ensuring scores from different forms of a test can be interpreted on the same scale. Differential Item Functioning (DIF) focuses on the comparability of item scores, rather than test scores. DIF is present when examinees from different language groups have different probabilities of answering an item correctly after conditioning on overall score (Zumbo, 1999). Chu and Kamata (2003) cautioned that DIF items may increase the errors of test equating or parameter estimates. When items show DIF, these items should be deleted before performing test equating. To meet both equating and DIF requirements, Chu and Kamata administered two different test forms linked by common-items to two groups of students.

In equating two different test forms administered in the same language, the anchor items are identical. However, in equating forms across languages (cross-lingual equating), it is unknown whether the anchor items are really identical, since they are translations of one another. In this study, we first evaluated the items for cross-lingual DIF and then choosed anchor items from the non-DIF items. Non-DIF items were treated as if they measured the same construct and have the same psychometrical characteristics. To evaluate cross-lingual DIF, we used a bilingual-group design to ensure that examinees proficient in both languages were tested in the two language versions of the test. In this single-group design, any difference in achievement across the two language versions of the test may be attributed to differences resulting from the translation/adaptation process.

3. Method

3.1 Instrument

The test studied was the 2005 Lower Secondary School Achievement mathematics test administered to 9th-grade students in Malaysia. Since 2003, this test has been delivered using dual-language test booklets where the items appear in English on one side of the booklet and in Malay on the other side. However, the Malaysian Ministry of Education is considering administering these exams only in English beginning in 2008. Both the English and Malay versions of the exam are designed to the same specifications and test the same content areas in mathematics.

The test consisted of 40 dichotomously scored multiple-choice items measuring topics such as numbers, algebra, measurement, geometry, and statistic. The data analyzed came from 505 students who were proficient in both the Malay and English languages. For this study, two separate booklets were prepared—one using only the English versions of the items, the other using only the Malay versions. The design for this study is a bilingual group design where each student took both language versions of the exam. To overcome a potential practice effect, the students were divided into two groups to counterbalance the order in which they took each language version of the exam.

3.2 Participants and design

The students participated in this study were students from four secondary schools in the state of Perak. For each school, intact class was used with students of high and average ability only involved. For every school, half of the students were administered the English version first and the other half administered the Malay version first. The final sample consisted of 505 students with the first group of 255 students took the English version mathematics test first while the second group of 250 took the Malay version of the test first. Three weeks later, each group was then given the other version. Both tests were administered during the last month of the school year.

The Malaysian education system adopted the bilingual program which stresses the academic use of both English and Malay languages. The students involved in this study received science and mathematics instruction in the English language when the Malaysian government implemented a policy that uses English as the language of instruction in the teaching and learning of science, mathematics, technical and technology subjects. The other subjects of Humanities and Arts continue to be in Malay language for all national secondary schools, while Malay, Mandarin or Tamil is also used as medium of instruction at the different ethnic primary schools.

The subjects in this study were bilingual students who are developing both oral and written communication skills in English and Malay at the same time. On the one hand, the students should be able to demonstrate what they know about mathematics on a test in the Malay language as it is the main language of instruction. On the other hand, the students' mathematics instruction was in English for the three prior years, and so their math terminology was probably more familiar in English. In addition, the students had been learning English as a school subject since grade one. Thus, the students were highly proficient in both languages and there may be advantages or disadvantages in testing them in either one of the two languages.

The bilingual group design is where a group of bilingual examinees, assumed to be equally proficient in both languages with respect to the construct being measured, is tested in the two language versions of the test (Sireci, 1997, 2005; Sireci & Berberoglu, 2000). The advantage of this design is that any differences in achievement between the two language versions can be attributed to differences the difficulty of the two versions, rather than to

differences in achievement between two different language groups.

4. Data analyses

4.1 DIF analyses

We evaluated the degree to which the items functioned similarly across their English and Malay versions by conducting Differential Item Functioning (DIF) analyses using the logistic regression procedure (Swaminathan & Rogers, 1990; Zumbo, 1999), which involves modeling the probability of answering an item correctly as a function of overall proficiency (total score on the test), group membership (in our case English or Malay version of the item) and the interaction of overall proficiency and group membership. The logistic regression equation is:

$$\ln\left[\frac{p_i}{(1-p_i)}\right] = b_0 + b_{tot} + b_{group} + b_{tot*group} \quad (1)$$

where p_i refers to the probability of responding to item i correctly, b_{tot} is the regression coefficient for the conditioning variable (e.g., total score), b_{group} is the regression coefficient for group membership, and $b_{tot*group}$ is the coefficient for the group-by-conditioning variable interaction.

Using logistic regression to detect DIF is a stepwise procedure. The first step enters the conditioning variable (total score) into the equation, to get a baseline proportion of variance accounted for. In the second step, the grouping variable (English or Malay item) is entered. In the third and last step, the interaction term (total score-by-group) is entered. The chi-square statistics (or regression coefficients) associated with steps 2 and 3 are used to determine statistically significant uniform or non-uniform DIF[1]. However, since the chi-square test is affected by sample size, effect sizes were also computed for each item. Items were classified as displaying negligible DIF, moderate DIF, or large DIF, according to the criteria established by Jodoin and Gierl (2001). The Jodoin-Gierl effect size criteria are based on the proportion of variance in item performance that can be accounted for by group membership (R^2) and are linked to the effect size classifications used by Educational Testing Service for the Mantel Haenszel statistic (Dorans & Holland, 1993). Using these classification rules, items were classified in the following manners:

(1) Negligible or A-level DIF: $R^2<0.035$,

(2) Moderate or B-level DIF: Null hypothesis rejected AND $0.035\leq R^2<0.070$,

(3) Large or C-level DIF: Null hypothesis rejected AND $R^2\geq 0.070$.

The second method to identify DIF items is carried out using computer program WINSTEPS (Linacre, 2003) based on Rasch Model. This model assumes that the item discrimination parameters are equal across the two groups being compared. By condition on person measure or group ability, item performance across groups is compared by estimating difficulty parameter for each group. Essentially, the difference in item difficulty parameter is assessed to account for group differences that cannot be explained by the test impact (Lord, 1980). Difference across two groups of examinees in item difficulty means that the item is more difficult for one group relative to the other group of examinees. For each group, WINSTEPS outputs estimates and standard errors for item difficulty. The DIF contrast which is the difference between the difficulty measures, and is a log-odds

[1] In the framework of IRT models, uniform DIF exists when the relative advantage across groups is uniform across the score scale. Nonuniform DIF occurs when the relative advantage for two groups differ across the score scale.

estimates equivalent to a Mantel-Haenszel DIF size. The procedure for quantifying DIF and testing for significance is based on the t-value computed with DIF contrast divided by the joint S.E. of the two DIF measures. A criterion t-value greater than 2.58 (p<0.01) is used to flag an item exhibits DIF.

4.2 Equating analyses

The equating analyses were conducted at two levels using classical test theory and Item Response Theory (IRT). For the first level, all 40 items on the test were included in the common subject design equating using Linking with Equivalent Group or Single Group Design (LEGS, Brennan, 2004) for classical approach, and BILOG-MG (Zimowski,2003) for the IRT method. LEGS links scores on two tests using various statistical methods including mean, linear, parallel-linear, and equipercentile methods with and without postsmoothing.

For the second level, the items identified as DIF were dropped before linking. The common items equating design is then conducted with Common Item Program for Equating (CIPE, Kolen, 2004) and BILOG-MG (Zimowski, 2003). For CIPE analysis, the non-DIF items are the common items (i.e., internal equating anchor) and the items flagged for DIF were the non-linking items.

The CIPE program implements the following equating methods: Tucker mean (TMEAN), Levine mean for internal common items (LMEAN), Braun/Holland mean (BMEAN), Tucker linear (TLIN), Levine linear for internal common items (LLIN), Braun/Holland linear (BLIN), unsmoothed frequency estimation equipercentile (UNSMOOTHED), and smoothed frequency estimation equipercentile, with up to 8 different degrees of cubic spline smoothing. In this study, only the analysis from the Tucker linear and equipercentile methods are reported. The CIPE calculates standard errors of equating for the Tucker linear, Levine linear, and unsmoothed equipercentile methods. The Tucker linear gives the smaller standard error of equating while the equipercentile method has the biggest standard error among the three methods.

We also conducted a separate common-item equivalent group equating using BILO-MG, again using the non-DIF items to anchor the scale. This was accomplished using concurrent calibration where the common (non-DIF) and unique (DIF) items were analyzed simultaneously using a 1-parameter logistic model[2].

4.3 Evaluation criteria

To evaluate the different approaches for adjusting the scores from one version of the exam to account for differences in test difficulty, we compared the raw score distributions for each form to each other and to the adjusted score distributions based on the equating analyses. We also compared the percentages of students who would pass the exam across the unadjusted and equated scores. The passing score for this exam was not released by the Malaysian government, but the national passing percentage was about 85% and so we chose a cut-score that resulted in an 85% pass rate for the Malay version of the test. We also compared the standard error of equating across the two classical approaches, and we conducted a likelihood ratio test to see if the IRT model that treated DIF items as non-equivalent provided a statistically significant improvement in fit to the data relative to the IRT model that treated all items as equivalent across the two languages (Thissen, Steinberg & Wainer, 1988).

5. Results

5.1 DIF analyses

[2] We did not consider more complex IRT models due to the relatively small sample sizes.

Table 1 summarizes the results from the logistic regression DIF analyses. Seven items were flagged for DIF using a criterion of statistical significance; however, all were classified as negligible or A-level DIF with R^2<0.035. Three of the items were easier in Malay, the other four were easier in English. The WINSTEPS analysis identified six items exhibiting DIF at t>2.58 (p<0.01). These six items were also flagged as DIF by logistic regression. WINSTEPS analysis is based on Rasch Model which allows for testing for uniform DIF only. The presence of uniform and/or nonuniform DIF can be tested simultaneously by logistic regression. The six DIF items flagged by both WINSTEPS and logistic regression indicate presence of uniform DIF while the seventh item identified by logistic regression could be nonuniform DIF.

Table 1 Summary of items flagged for DIF

Item No.	b_R English	b_F Malay	T	Flagged in LR	R^2 effect size	Favors
13	1.43	0.96	2.99*	Yes	0.010	M
18	0.07	0.55	-2.78*	Yes	0.012	E
20	0.12	0.58	-2.72*	Yes	0.009	E
23	-0.18	0.30	-2.65*	Yes	0.012	E
25	1.13	1.70	-3.67*	Yes	0.012	E
32	0.33	-0.61	4.94*	Yes	0.035	M
33	2.49	2.12	2.32	Yes	0.010	M

Notes: All items were flagged for DIF at p<0.01 using logistic regression. Items marked with * were also statistically significant at p<0.01 using Winsteps.

5.2 Equating analyses

The first level equating results were obtained by including all items in the test, while the second level was conducted with the exclusion of the seven DIF items flagged by logistic regression. The results from both approaches are summarized in Table 2.

Table 2 English math score using linear and unsmoothed equipercentile equivalents

Malay math score	Equating with DIF items		Equating without DIF items	
	L_Y (SE)	e_Y (SE)	L_Y (SE)	e_Y (SE)
10	8.1 (0.36)	12.0 (0.86)	10.2 (0.27)	9.7
20	18.6 (0.24)	17.8 (0.33)	20.1 (0.17)	20.6 (0.30)
22	20.6 (0.21)	20.0 (0.36)	22.1 (0.15)	22.1 (0.29)
30	29.0 (0.16)	29.6 (0.06)	30.0 (0.08)	29.7 (0.26)
40	39.5 (0.06)	39.7 (0.10)	39.9 (0.09)	40.0 (0.00)

Notes: L_Y – linar equivalents; e_Y – equipercentile equivalents; SE – standard error.

Using the lowest quartile Malay math score of 10, the equivalent English score is 8.1 for the linear equivalent before deleting DIF items. Results from the equipercentile are not considered as S.E. is large (0.86) which is probably due to the small number in the low achieving group. When equating excluding the DIF items, the performance of the low achieving students improves to 10.2 and is almost the same as performance in Malay math test.

Using the maximum and upper quartile score of 30 and 40, the English math score equivalent is almost the same as the Malay math score without DIF items (30 and 39.9 for the linear equivalent and 29.7 and 40.0 for the equipercentile equivalent). A special comparison is considered for the passing score of 22 (85% pass), the adjustment is adding 1.4 point with the linear equating method and adding 2 points with the equipercentile method for including DIF items. Similarly, equating results improved with DIF items excluded which require an adjustment of only 0.1 point and yet with smaller S.E.

Table 3 shows the equating results using IRT method. Equating including all items showed that the score difference is small for higher score. For the high achieving group, the English math score equivalent is almost the same as the Malay math score. For the passing score of 22, the adjustment is again 1.6 points with equating inclusive of DIF items.

Table 3 English math score equivalent using IRT estimated true score

Equating with DIF items			Equating without DIF items		
Malay Math score	θ-equivalent	English math score equivalent	Malay Math score	θ-equivalent	English math score equivalent
40	1.60	39.8	40	1.64	40
30	-0.44	29.0	30	-0.42	30
22	-1.31	20.4	22	-1.29	22
20	-1.33	18.5	20	-1.33	20

Interestingly, when equating without the DIF items, the Malay math scores are equivalent to the English math scores at 20, 22, 30 and 40. Thus, the passing score of 22 seems equivalent across the two language versions after equating, regardless of how the items flagged for DIF items were treated.

The correlation between the theta values computed from the BILOG analysis was computed. The correlation relationship between math abilities in English and Malay was essentially the same with and without the DIF items included in the scaling (0.607 with DIF items and 0.609 without DIF items). The strength of the relationship indicates that language may be an issue in assessing the math ability as the same examinee took the two language versions of the math test.

In addition to correlation, the relationship between the achievement in Malay version and English version of the math test is display in Figure 1. The plot shows the score distribution between the Malay math scores in relation to the English math score. Based on the cutoff score of 22, 81% of the examinees pass the test whether it is in Malay or the English version. 2.8% can only pass the test in English, while 8.5% pass the math test if administered in Malay. This group of students may be misclassified if the math test were administered in English. The rest of the 7.7% fail the math test whether it is in Malay or English. This gives a decision consistency of 0.887. The Kappa statistic computed, $\kappa=(0.89-0.65)/(1-0.65)=0.69$. These results are summarized in Table 4, where it can be seen that 89.5% pass the test in Malay, while only 83.8% pass the test in English.

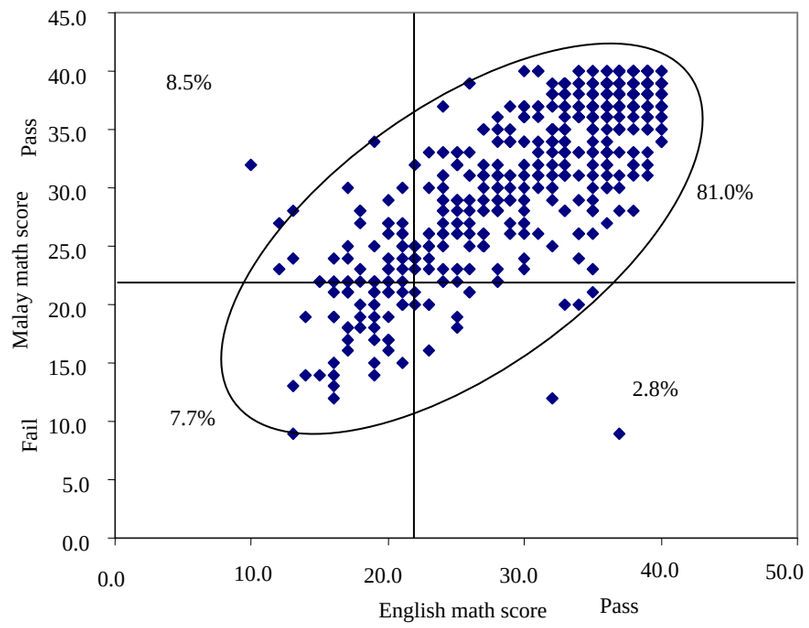

Figure 1 Malay math score versus English math score

Table 4 Percentage pass/fail math test in English and in Malay

		Math test in Malay		
		Fail	Pass	Total
Math test in English	Pass	2.8% (14)	81.0% (408)	83.8%
	Fail	7.7% (39)	8.5% (43)	16.2%
	Total	10.5%	89.5%	100%

Table 5 Percentage pass/fail in English and Malay math test with adjusted score

		Math test in Malay		
		Fail	Pass	Total
Math test in English	Pass	4.8% (24)	84.7% (427)	89.5%
	Fail	5.7% (29)	4.8% (24)	10.5%
	Total	10.5%	89.5%	100%

Using the equating results, the equivalent passing score for the English math test is adjusted for 2 points (i.e., the passing score on the English version is set at 20). After this adjustment, 84.7% of the examinees pass the math test both in English and Malay. About 4.8% pass the test in English, 4.8% pass if it is in Malay, and 5.7% continue to fail whether it is given in English or Malay. The decision consistency for passing or failing is 0.91, which is slightly higher than before equating. The Kappa statistic computed has a small increase, with $\kappa=(0.91-0.73)/(1-0.73)=0.67$. A scatterplot of the scores after equating is presented in Figure 2.

With the score adjusted after equating, the percentage pass for the math test in the two language version is the same, that is, 89.5%.

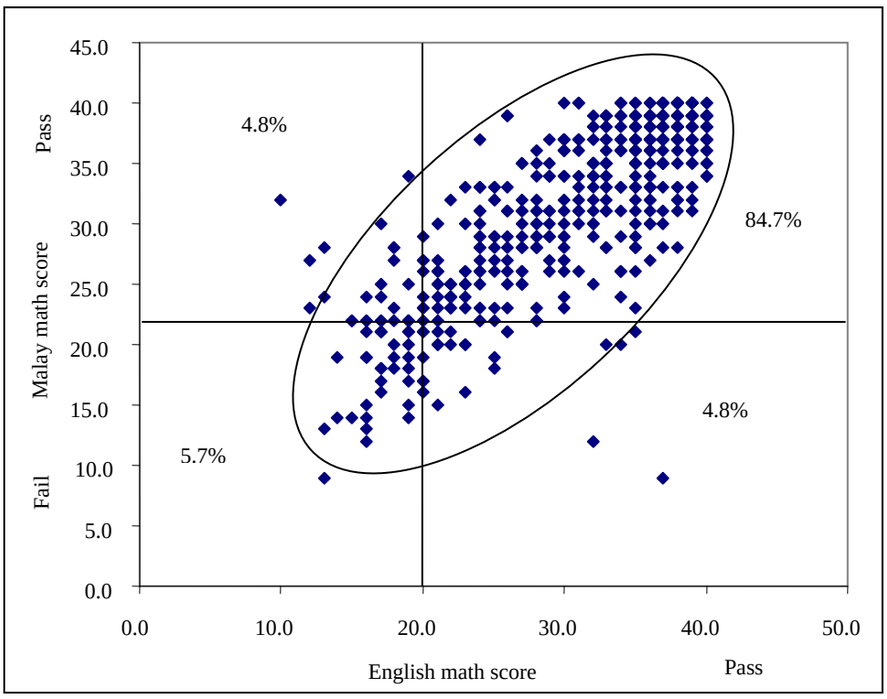

Figure 2 Malay math score versus English math score after equating

6. Discussion and conclusion

Administering tests to students who operate in different languages is a difficult endeavor. As the results of this study show, when students can theoretically take a test in two languages, the language in which they take it can have an impact on their performance, if differences in difficulty across the test forms are not accounted for.

Our results show that deleting DIF items greatly improved equivalence of the two language versions of the math test based on both the classical and IRT methods. Both linear and unsmoothed equipercentile methods gave similar results when equating without DIF items, which indicates that the two language version of math test is similar. A bigger adjustment is necessary for comparison if equating is performed without deleting DIF items. Thus, when equating test forms, it is important that DIF items should be deleted.

There are bilingual students who pass only the Malay version of the math test. Assessing their mathematics proficiency in English may provide inaccurate information in making decisions about their math achievement. Instead, a language accommodation (i.e., administering the test in Malay) may be necessary for valid assessment. A similar argument can be made for students who only passed the Malay version. Clearly, evaluation of the effect of language of administration on students' performance is needed before interpreting students' scores whenever the language of the test is not commensurate with the language of instruction or with the student's native language proficiency.

From a methodological perspective, analyses based on bilingual examinees have several advantages. First, there is no sample size issue as is seen in some cross-lingual studies where one language group dwarfs the other. Second, the examinees are truly equivalent with respect to the construct measured, and differ only in their ability

to access the construct in either language. The fact that each examinee is probably stronger in one of the two languages prohibits us from using bilinguals to equate test forms in a strict sense, but the results give us important information regarding comparability of the forms, particularly when the bilinguals are sufficiently proficient in each language, as they were in this study.

References:

Brennan, R. (2004). *Linking with equivalent group or single group design (LEGS) (version 2.0)*. University of Iowa: Center for Advanced Studies in Measurement and Assessment (CASMA).

Chu, K. L. & Kamata, A. (2003). Test equating with the presence of DIF. Paper presented at *the Annual Meeting of American Educational Research Association*, April 2003, Chicago.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In: P. W. Holland & H. Wainer. (Eds.). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum, 35-66.

Gao, H. (2004). *The effect of different anchor tests on the accuracy of test equating for test adaptation*. (Doctoral dissertation, Ohio University)

Jodoin, M. G. & Gierl, M. J . (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329-349.

Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. In: R.K. Hambleton, P. Merenda & C. Spielberger. (Eds.). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum, 3-38.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating: Methods and practices*. New York: Springer-Verlag.

Kolen, M. J. (2004). *CIPE: Common item program for equating (CIPE) (version 2.0)*. University of Iowa: Center for Advanced Studies in Measurement and Assessment (CASMA).

Linacre, J. M. (2003). *WINSTEPS: Rasch-model computer programs*. Chicago: Winsteps.com.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Rapp, J. & Allalouf, A. (2003). Evaluating cross-lingual equating. *International Journal of Testing, 3*(2), 101-117.

Sireci, S. G. (1997). Problems and issues in linking assessment across languages. *Educational Measurement: Issues and Practice, 16*(1), 12-19.

Sireci, S. G. & Berberolu, G. (2000). Using bilinguals to evaluate translated assessment questions. *Applied Measurement in Education, 13*(3), 229-248.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In: R. K. Hambleton, P. F. Merenda & C. D. Spielberger. (Eds.). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, 117-138.

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*. Hillsdale, NJ: Erlbaun, 147-169.

Zimowski, M., Muraki, E., Mislevy, R. & Bock D. (2003). *Bilog-MG (computer software)*. Mooresville, IN: Scientific Software International.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

(Edited by Max and Nydia)