

Using Bilingual Students to Link and Evaluate Different Language Versions of an Exam

Ong Saw Lan

Malaysia Science University

Stephen G. Sireci

University of Massachusetts Amherst

## Using Bilingual Students to Link and Evaluate Different Language Versions of an Exam

### 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 between the different language versions of the test be 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 9<sup>th</sup>-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 evaluates the utility of gathering data from bilingual students for the purposes of evaluating differences in difficulty across test forms due to translation.

### Test Equating and DIF

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 testing forms if it is two versions of a test with the same content. In the case of test adaptation or translation, statistical procedure utilized for adapted exams to adjust test scores on different language versions of the same exam so that scores 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 CTT method, score correspondence of test is established by setting characteristics of the

score distribution equal for a specified group of examinees (Kolen & Brennan, 2004). Three often used CTT methods are mean equating, linear equating, and equipercentile equating.

DIF is present when examinees from different language groups have different probability of answering an item correctly after conditioning on overall score (Zumbo, 1999). Chu & Kamata (2003) have the opinion that DIF items increase the errors of test equating or parameter estimates. When items show DIF, these items should be deleted before performing test equating. In test equating, scores of multiple test forms are placed onto a common scale. On the other hand, the design of DIF analysis requires different groups to take the same test form so that group differences will show on test items. In order to meet both equating and DIF requirements, Chu & Kamata administered two different test forms linked by common-items to two groups of students.

In same-language equating the anchor items are identical. For the cross-lingual equating in this study the anchor items are chosen from the non DIF items. These items are treated as if they measure the same construct and have the same psychometrical characteristics. The bilingual design group is adopted to assure that examinees proficient in both languages are tested in the two language versions of the test. If there is a difference in achievement between the two language versions, it is probably attributed to differences in the difficulty of the two versions.

## Method

### *Instrument and Participants*

The test studied was the 2005 Lower Secondary School Achievement mathematics test administered to 9<sup>th</sup>-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. The two tests are identical but in two different languages, practice effect may be a problem. To overcome this, the students were divided into two groups. The first group of 255 students took the English version mathematics test while the second group of 250 took the Malay version of the test. Three weeks later, the first group was then given the Malay version while the second group took the English 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 have 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 are bilingual students who are developing both oral and written communication skills in two languages at the same time. The students are able to demonstrate what they know about mathematics on a test in the Malay language as it is the main language of instruction. However, mathematics instruction has been in English for the last three years in addition to learning English as a school subject since grade one, testing in Malay would yield misleading results due to unfamiliarity with mathematics terminologies. Students may better demonstrate the mathematical knowledge that they have in English, the language of instruction.

The bilingual group design (Sireci, 1997) 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. 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.

### *Data Analyses*

#### *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. 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 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 manner:

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

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

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 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 ( $p < .01$ ) is used to flag an item exhibits DIF.

### *Equating Analyses*

The equating analyses were conducted at two levels using classical test theory and item response theory (IRT). For the first level, all the 40 items in 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 seven items identified as DIF are 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 33

translated non-DIF items are the common items and the seven flagged DIF items as the non-linking items conducted with internal equating.

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 are used. 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.

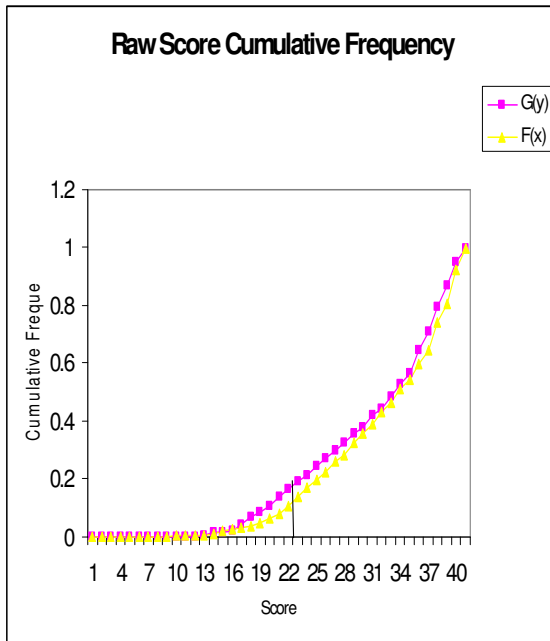
Common items equivalent group equating is repeated using BILO-MG. The common items are the 33 DIF free items for the two test forms (English and Malay math test), and the seven DIF items are the unique items. Concurrent calibration where the common and unique items are analyzed simultaneously is used in the 1-parameter logistic model.

### *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.

### Results

Figure 1 shows the score distribution with all the 40 test items used for computation of score. For the score range between 17 – 29, the score distribution for English is higher than the Malay and slightly higher for the range between 35 – 38. The score distribution removing the seven DIF items is as shown in Figure 2, indicating a smaller difference between the English and the Malay version score distribution.



Note: G(y) Math test in English,  
F(x) Math test in Malay

Figure 1

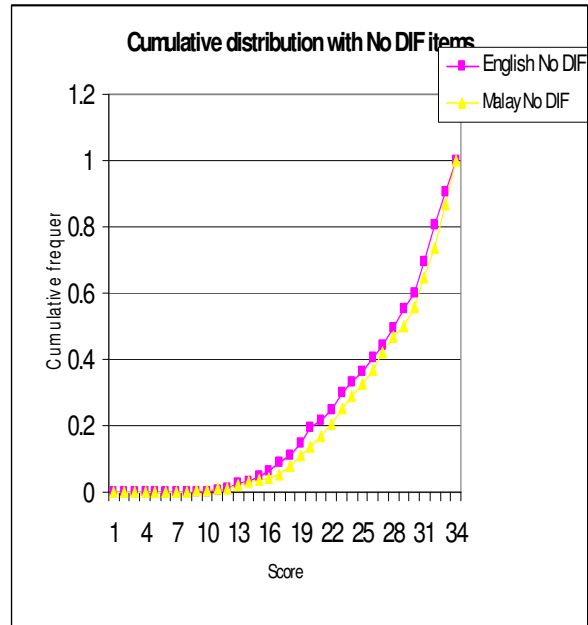


Figure 2

Table 1 shows that results from the logistic regression flagged seven items demonstrating DIF. These are item number 13, 18, 20, 23, 25, 32, and 33. All these items are classified as negligible or A-level DIF with  $R^2 < 0.035$ . The WINSTEPS analysis identified six items exhibiting DIF at  $t > 2.58$  ( $p < .01$ ). These six items were also flagged as DIF by logistic regression.

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	.96	2.99*	Yes	.010	M
18	.07	.55	-2.78*	Yes	.012	E
20	.12	.58	-2.72*	Yes	.009	E
23	-.18	.30	-2.65*	Yes	.012	E
25	1.13	1.70	-3.67*	Yes	.012	E
32	.33	-.61	4.94*	Yes	.035	M
33	2.49	2.12	2.32	Yes	.010	M

Note: All items were flagged for DIF at  $p < .01$  using logistic regression. Items marked with \* were also statistically significant at  $p < .01$  using Winsteps.

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 obtain is shown in Table 2.

Table 2  
Rounded English math score using Linear ( $L_Y$ ) and Unsmoothed Equipercentile ( $e_Y$ )  
Equivalents

Malay Math Score	Equating with DIF items		Equating without DIF items	
	$L_Y$ (S.E.)	$e_Y$ (S.E.)	$L_Y$ (s.e.)	$e_Y$ (s.e.)
10	8.1 (.36)	12.0 (.86)	10.2 (.27)	9.7
20	18.6 (.24)	17.8 (.33)	20.1 (.17)	20.6 (.30)
22	20.6 (.21)	20.0 (.36)	22.1 (.15)	22.1 (.29)
30	29.0 (.16)	29.6 (.06)	30.0 (.08)	29.7 (.26)
40	39.5 (.06)	39.7 (.10)	39.9 (.09)	40.0 (.00)

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 is not considered as S.E. is large (.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 improve 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 1.4 point with the linear equating method and 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  
English Math score equivalent using IRT Estimated True score

Equating with DIF items			Equating without DIF items		
Malay Math score	$\theta$ -equivalent	English math score equivalent	Malay Math score	$\theta$ -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

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 point with equating inclusive of DIF items using the 1-parameter IRT model.

When equating without the DIF items, the Malay math scores are equivalent to the English math scores at 20, 22, 30, and 40.

The correlation between the theta values computed from the BILOG analysis was computed. The relationship between math abilities in English and Malay is medium; 0.607 with DIF items and 0.609 without DIF items. The strength of the relationship indicates that language is an issue in assessing the math ability as the same examinee took the two language version 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 is 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 .887. The Kappa statistic computed,  $\kappa = (.89-.65)/(1-.65) = .69$ .

Table 4 indicated that 89.5% pass the math test in Malay, while only 83.8% pass the math test in English.

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%

Using the equating results, the equivalent passing score for the English math test is adjusted for 2 points. The passing score is set at 20. Now 84.7% of the examinee pass the math test both in English and Malay, which is slightly higher than before equating. 4.8% pass the test only if it is in English and another 4.8% pass if it is in Malay. 5.7% continue to fail whether it is given in English or Malay. The decision consistency for passing or failing is .91, which is slightly higher than before equating. The Kappa statistic computed has a small increase, with  $\kappa = (.91-.73)/(1-.73) = .67$

Table 5:  
Percentage pass/fail in English and Malay math test with adjusted score <sup>84.7</sup>

		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%

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%.

### Discussion and Conclusion

The 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 unsmooth equipercentile methods gave similar score when equating without DIF items. This 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 math test in English only version may provide inaccurate information in making decision about students' math achievement. Language accommodation may be necessary for valid assessment.

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)* [computer software]. 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* (pp. 35-66). Hillsdale, NJ: Erlbaum.
- Gao, H. (2004). *The Effect of Different Anchor Tests on the Accuracy of Test Equating for Test Adaptation*. Ph.D. 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* (pp. 3-38). Hillsdale, NJ: Lawrence Erlbaum.
- Kolen, M.J. & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2<sup>nd</sup> ed.)*. Springer. New York.
- Kolen, M. J. (2004). *CIPE: Common Item Program for Equating (CIPE) (version 2.0)* [computer software]. University of Iowa: Center for Advanced Studies in Measurement and Assessment (CASMA).
- Linacre, J.M. (2003). WINSTEPS: Rasch-Model: Computer Programs.
- 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.
- Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- 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.