Asia Pacific Journal of Educators and Education, Vol. 29, 1–18, 2014

THE VALIDITY OF ADMINISTERING BILINGUAL MATHEMATICS TEST AMONG MALAYSIAN BILINGUAL STUDENTS USING DIFFERENTIAL ITEM FUNCTIONING (DIF)

S. Kanageswari Suppiah Shanmugam[1*] and Ong Saw Lan[2]

[1]SEAMEO Regional Centre for Education in Science and Mathematics,
SEAMEO RECSAM Pulau Pinang
[2]School of Educational Studies, Universiti Sains Malaysia, 11800 Pulau Pinang

*Corresponding author: kanageswari@recsam.edu.my

Abstract: Language dissimilarity is a liability in translated tests which can result in construct non-equivalence as the construct under measure in one test cannot be generalised to the construct of another test in a different language. This study examines the validity of using bilingual test booklet to assess Form Two Malaysian students' mathematical achievement. The equivalence of the two test booklets were inspected using differential item functioning methodology. The sample consists of 4,768 Form Two students from 34 schools in two states in Malaysia. The instruments used were bilingual mathematics test (English and Malay languages) and English-only mathematics test. Both tests contain the same 40 multiple-choice items where 20 are word-problem and another 20 are computation items that were obtained from 1999 and 2003 TIMSS released items. Random group design with spiral administration was employed. Differential Item Functioning (DIF) analyses performed were based on the Mantel Haenszel method, Multidimensional model and IRT 1-PL model. Only one common DIF item was flagged which made up 2.5% of the test items. Therefore, the two test booklets are equivalent. Test validity was not compromised with the use of the bilingual test. This study has important implication on Mathematics test item writers who need to remove language redundancy when developing test items and constructing tests.

Keywords: bilingual testing, differential item functioning, validity

INTRODUCTION

The English language facilitates the acquisition of scientific knowledge (Ainan, 2003; Ain Nadzimah, & Chan, 2003). A good command in the English language will assist Malaysian students to access articles which are commonly published in English (Ainan, 2003; Pandian & Ramaiah, 2004; Ministry of Education, 2004; Yoong, 2005) and since these texts are in English, a new policy that reshuffled the language policy of the Malaysian national education system was announced in 2002. The English-Medium Instruction Policy states that as of 2003, the English language would be the language of instruction in the teaching of

Mathematics and Science subjects for the Year One, Form One and Lower Six students while the languages of assessment would be in the Malay language and English language (Surat Pekeliling Ikhtisas Bil 11/2002).

Since English is Malaysian students second or third language, they lacked the level of English proficiency required to fairly demonstrate their mathematical skills especially in word problems mathematics items (Fatimah & Zarina, 2004). Since all assessments measure language skills (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999), for limited English proficiency (LEP) students, when testing is in English, language becomes a measure of that test construct (August & McArthur, 1994). As such, the test scores may not reflect the latent trait that is being assessed and this may compromise reliability, validity and fairness – the fundamentals of any assessment (AERA et al., 1999; Educational Testing Service [ETS], 2002).

To alleviate the challenges posed by language, assessing students in their dominant language is an important consideration. A bilingual test can address this linguistics complication as it makes available test items in students' dominant language (AERA et al., 1999). Bilingual test gives provision for the original test items to be translated into the students' dominant language that they are more likely to be proficient in and as such, it removes the unnecessary language barrier (AERA et al., 1999). Among Malaysian students who are non-native speakers of English, bilingual assessment is a viable test accomodation (Ainan, 2003; Ministry of Education, 2004) as it is able to cushion the sudden implementation of English as the language of instruction.

PROBLEM STATEMENT

When abilities other than the intended target ability are assessed and are reflected in the test score, test validity is questioned. This is according to Shealy and Stout (1993), the main construct of the test that is intended to be measured (primary dimension) is influenced by the ability that the test does not intend to measure (secondary dimension). Therefore, the abilities other than the target ability that the test intends to measure are likely to influence students' test performance and is reflected in their score.

Differential item functioning (DIF) occurs when a secondary dimension is added to the test construct. Therefore, this secondary dimension (Shealy & Stout, 1993) causing items function differently for different groups of students with the same ability but the probability of producing a correct response differs due to the interference of the secondary dimension. This results in DIF (Bechger, 2006).

This secondary dimension is auxiliary if it is intentionally assessed and becomes benign DIF if it is an unintentional construct.

Gierl and Khaliq (2001) detected DIF items across different language versions of the test as translation does not ensure psychometric equivalence between the tests. In a Mathematics test that was translated from English to French, Gierl, Rogers and Klinger (1999) suggested that two of the seven Mathematics items that were detected as DIF were caused by translation. The results indicated that DIF analyses helped to identify items that are flawed due to translation or other possible causes (Allalouf, Hambleton & Sireci, 1999). DIF items can then be refined or removed to eliminate construct-irrelevant variance from the language barrier that threatens construct validity (Messick, 1995).

AIM

The main aim of this research is to detect DIF items in a bilingual test with reference to an English-only test. The objective of this research is to compare the number of DIF items and the type of DIF items that are detected in a bilingual test with the English-only test.

RESEARCH SIGNIFICANCE

As test scores reflect students' ability, this research is of great significance as it sheds light on whether test items in different languages demonstrate equivalence and enable correct decision to be made based on different language versions of test.

RESEARCH LIMITATIONS

During the data collection, information on the students' language background, particularly the language of instruction used at the primary school (which can be in either Malay, Chinese or Tamil language) was not gathered. In retrospect, this information is vital to see whether the Malay-English bilingual test is useful to all Malaysian students irrespective of the language of instruction while they were in the primary school or is affected by the language of instruction.

DIFFERENTIAL ITEM FUNCTIONING (DIF)

Items that function differently for different groups of people can be identified through DIF analysis as it identifies items that display psychometric differences which signals that the items are functioning differently for two groups matched by the measured construct (Bechger, 2006). According to Gierl and Khaliq (2001), DIF items occur across different language test forms because when translating tests items does not ensure psychometric equivalence between the different test languages. To avoid false identification of DIF items, Ercikan, Gierl, McCreith, Puhan and Koh (2004) proposed the use of at least two methods to flag DIF items.

In their study, Ercikan et al. (2004) showed that 18% to 31% of the Mathematics item in English and French versions exhibited DIF. Even for items in international testing for TIMSS, Ercikan and Koh (2005) found the English and French items displayed DIF. The English and French versions were administered in Canada with 22 or 14% of the items displaying DIF. However, there were a higher number of DIF items detected in the French items that were administered in France compared to the French items administered in England and United States of America. In comparison, in England-France sample, 61 items or 39% displayed DIF while in the United States-France group, 91 or 59% of the items displayed DIF. Similar findings were made by Allalouf, Hambleton and Sireci (1999). From 125 items, they detected 42 DIF items on the Israeli Psychometric Entrance Test that has been translated from Hebrew to Russian when compared between the Hebrew and Russian speaking samples.

Gierl and Khaliq (2001) identified translation as a cause of DIF when words that affect meaning are either omitted or added. As a result, the difficulty of the words which are either inherent in the language or are not inherent in the language are altered across languages. In addition, difficulty in the item format is another possible cause of DIF. In addition, changes in words or sentences, in content, in the test format and culturally-related issues are also sources of DIF since they affect the test difficulty (Sireci & Allalouf, 2003).

DIF could be the result of differences inherent in the item characteristics or in the students' ability. If the difference is due to the nature of the item, then the source of DIF needs to be determined. (Elosua & Lopez-Jauregui, 2007). Only then the expert can determine whether the performance difference is due to construct relevant or construct irrelevant variance. According to Hofstetter (2003), cultural and linguistic factors which include the written and spoken form of language in the social and academic setting can cause bias.

It is crucial to note that not all DIF items may be bias. Statistical analyses are useful only to flag DIF items as they cannot be used to determine whether the DIF is the result of test bias or test impact (Boughton, Gierl & Khaliq, 2000). Impact occurs when the differences in test performance is caused by the differences in the ability of the two different groups (Gierl, 2004). However, the presence of DIF items does not necessarily indicate a test weakness as items may consistently function differently for groups of students with special personal traits which points to the emergence of multidimensionality (AERA et al., 1999).

DIF Comparison Across Classical Test Theory and Item Response Theory

When students are being assessed in a language other than their first language, their scores may not accurately reflect their true ability as language is also part of the construct being measured. Due to the language factor, a secondary dimension is added to the test construct which may be irrelevant (Short & Spanos, 1989) and compromises the primary dimension that the test intends to measure. According to Shealy and Stout (1993), this ability that the test does not intend to measure is a nuisance determinant that violates test validity and produces adverse DIF.

Three methods of DIF were used in this study which are the Mantel Haenszel method that are based on Classical Test Theory (CTT) and IRT 1-PL and, the multidimensional IRT model. For the Mantel Haenszel method, WINSTEP that is based on IRT 1-PL and LERTAPS that is based on CTT were used while for the multidimensional IRT model, SIBTEST was used.

Based on CTT, DIF analysis is conducted based on *MH D-DIF* and the *p*-value of less or equals to 0.05 which indicates significance. *MH D-DIF* represents the Mantel Haenszel size (Nelson, 2001b).

Based on IRT, DIF items are flagged by using the Mantel Haenszel method that provides DIF effect (Fidalgo, Ferreres, & Muniz, 2004). The Mantel Haenszel method is based on logit-linear theory where DIF uses log-odd estimators (Linacre, 2008b).

For the multidimensional IRT model, SIBTEST measures the size of DIF and classifies the items as having negligible, moderate or large DIF by using effect size (Gierl & Bisanz, 2006). The beta estimate describes the effect size of the DIF (Stoneberg, 2004).

RESEARCH DESIGN

The research adopted the random group equivalent design with spiral administration which involves assigning different test booklets to students within

the same stipulated testing time (Forsyth, Hambleton, Linn, Mislevy & Yen, 1996). For each school six classes were selected. In each class, student was assigned either the English-only or the bilingual test. By alternating each test booklet, all the students in the class sat for one Mathematics test and this approach allowed the test booklet effect to be controlled (Duncan, del Río Parent, Chen, Ferrara, Johnson, Oppler, & Shieh, 2005).

Students' mathematical achievement was measured by using raw scores obtained in the Mathematics test. The group of students who answered the English-only test is the control group while students who answered the bilingual test is the focal group. To compare for the effect of the bilingual test booklets, DIF analyses were performed. If the percentage of the DIF items is small, then the availability of the additional language has not altered the test. If the percentage of the DIF items is significantly large, then what is measured by the test may has been altered due to the emergence of a secondary dimension which probably is caused by the language.

The Mathematics Test

The first stage involves identifying and selecting items that conformed to the Form One and Form Two Malaysian Mathematics Curriculum which were defined in the learning outcomes of the Curriculum Specifications Mathematics Form One (Curriculum Development Centre, 2003a) and Curriculum Specifications Mathematics Form Two (Curriculum Development Centre, 2003b) from the released grade eight TIMSS 1999 and 2003 Mathematics items. By using the expert judgment of three Mathematics teachers, a set of 62 items were identified. These items were pilot tested. After which 40 items were selected where content validity overruled statistical properties as statistical features only served as a guide (Henrysson, 1971) and that the test content must be well-represented (Tinkelman, 1971). These items were translated to the Malay version.

The English-only version consisted of two parts. The first part included the personal particulars of the subjects. The second part consisted of 40 multiple choice questions. These items were composed of either three or four distracters with one correct response.

The bilingual (Malay and English) test consisted of three parts. The first part addressed the subjects' personal particulars. The second part consisted of the same 40 multiple-choice items that appeared in the English-only version, with an added Malay language version. For each item, the Malay language version appeared immediately after the English form, in a square parenthesis using bold italic print of the same font size. The third part addressed the usefulness of the bilingual version by eliciting responses from students on the extent of using the

Malay language version to understand the items. Specifically, they were requested to write down the item numbers for items that they had difficulty understanding in English and resorted to using the Malay version to understand them.

Subjects

The subjects were Form Two students. The student sample design employed was a two-stage cluster sampling of schools at the first stage and classroom at the second stage. Cluster sampling was adopted at each stage.

The number of schools selected were 34 schools where 17 schools were from the Penang island, 12 schools from Penang mainland and five schools from Perak. Purposive sampling was used as only schools where the teachers had completed all the topics assessed in this test which is until Chapter 14 of the Form Two syllabus before the first week of October were selected. In each school, six Form Two classes were selected and within each class, all the students were selected. 12 schools had less than six classes and as such all the classes were selected. All together, the study used 4,768 students with 2,399 students who sat for bilingual test and 2,369 for the English-only test.

Data Analyses

The correct response was scored '1' while an incorrect response or unanswered items was scored '0'. DIF analyses was determined by using Mantel Haenszel method based on CTT and IRT and, the multidimensional model were employed. To conduct the DIF analyses based on IRT, WINSTEPS Version 3.67.0 (Linacre, 2008a) was used and for DIF analyses based on CTT, LERTAP Version 5 (Nelson, 2001a) was used. DIF analyses based on the multidimensional model was conducted by using SIBTEST (Stout, 2005).

When performing the DIF analyses, students were divided into the focal and reference groups where the students who sat for the bilingual test is the focal group while examinees who answered in the English-only test was treated as the reference group. To identify DIF items, classification is based on the DIF method that is employed as the magnitude of DIF item determines whether the item is displaying negligible, moderate or large DIF.

Mantel Haenszel Method

The output generated by LERTAP l classifies the DIF items as displaying negligible, moderate or large DIF based on the magnitude of *MH D-DIF* that represents the Mantel Haenszel size. When *MH D-DIF* is between –1 and 1, the

DIF item is negligible. If the value is between –1.5 and 1.5 then moderate DIF is flagged and any value outside this range will flag large DIF. All values are statistically significance at $p < .05$. *MH D-DIF* that is positive favours focal group while negative *MH D-DIF* favours reference group (Nelson, 2001b).

Mantel Haenszel Method using IRT 1-PL

Using WINSTEPS, the students' ability measures or θ are sliced into strata. The DIF contrast is the difference between the item difficulty, *b* for the two groups. For an item to be flagged as DIF, items can be classified as exhibiting negligible, moderate or large DIF based on the following criteria for the DIF size (Zwick, Thayer, & Lewis, 1999).

C = moderate to large $|DIF| \geq 0.64$
B = slight to moderate $|DIF| \geq 0.43$
A = negligible $|DIF| < 0.43$

For statistically significance DIF on an item, $p < .05$. Positive value favours the focal group while negative value favours the reference group (Linacre, 2008b).

The Multidimensional Model

The multidimensional model was also used to detect DIF by using SIBTEST. To run SIBTEST, the items were divided into two groups which are the suspected subtest and the matching subtest. The computation items that displayed primary dimension was classified as matching subtest while the word problem items that displayed the primary and secondary dimensions were put into the suspected subtest. Each item is tested against all other items.

An item is flagged as DIF if the *p-value* is less than 0.05. SIBTEST measures the size of DIF and classifies the items as having negligible, moderate or large DIF by using effect size (β_{UNI}). β_{UNI} is estimated as $\int B(\theta) f_F(\theta) d\theta$ where $B(\theta) = P(\theta, R) - P(\theta, F)$ which is the difference between the probabilities of the examinees' correct response from the focal and reference groups, $f_F(\theta)$ represents the density function for θ in the focal group and *d* being the width of the scaling interval. β_{UNI} is computed by integrating the product of $B(\theta)$ and $f_F(\theta)$ over θ. SIBTEST is used to calculate β_{UNI} by using the test statistics,

$$SIB = \frac{\beta_{UNI}}{\sigma(\beta uni)}$$

where $\sigma\beta_{UNI}$ is the estimated standard error of β_{UNI} (Gierl & Bisanz, 2006).

Two important statistical indices are the *p-value* that provides an estimate to the probability that the observed difference is due to chance and the beta estimate that describes the DIF effect size (Stoneberg, 2004). The magnitude of the DIF items is based on the beta estimate. According to Roussos and Stout (1996), negligible DIF occurs for beta estimate value that is less than 0.059 while moderate DIF exists for beta estimate value that is between 0.059 and 0.088. Large DIF is flagged by beta estimate value that exceeds 0.088. A positive value of β_{UNI} indicate that the flagged DIF item is against the focal group while a negative β_{UNI} is against the reference group.

RESULTS

Differential Item Functioning Analyses for the Two Tests

DIF analyses were conducted by using IRT, CTT and the multidimensional model. Based on IRT analysis as shown in Table 1, there were nine items flagged as DIF where five are computation items and four word problem items. Only one word problem item exhibited moderate DIF while the rest were categorized as negligible DIF. The four word problem DIF items are Items number 10, 15, 16 and 40 and the five computation items are Items 4, 5, 25, 34 and 38. Item 40 recorded Mantel Haenszel chi-square size of 0.45 (more than the cutoff value of 0.43), indicating slight to moderate DIF according to the criteria suggested by Zwick et al. (1999).

Table 1. DIF Items based on IRT

Item Number	p	Mantel Haenszel Size	DIF
C4	0.037	–0.17	Negligible
C5	0.013	–0.17	Negligible
C25	0.001	–0.23	Negligible
C34	0.023	–0.18	Negligible
C38	0.009	0.18	Negligible
W10	0.001	–0.24	Negligible
W15	0.010	0.19	Negligible
W16	0.021	0.15	Negligible
W40	0.000	0.45	Moderate

Four of the DIF items (Item W15, W16, W40 & C38) have positive Mantel Haenszel chi-square value indicate that these items favour the focal group, that is, students who answered the bilingual test. As three of the items are word problem items, this may due to the availability of the additional Malay version of the test that helped the students to comprehend the word problem items better compared to the English-only test items.

The DIF analyses form the CTT analyses identified the same nine items as exhibiting DIF. Again, only one item, Item W40 was flagged as moderate DIF as indicated in Table 2.

Table 2. DIF Items based on CTT

Item Number	p	Mantel Haenszel Size	DIF
C4	.04	0.42	Negligible
C5	.01	0.40	Negligible
C25	.00	0.54	Negligible
C34	.03	0.40	Negligible
C38	.01	0.42	Negligible
W10	.00	0.54	Negligible
W15	.01	–0.42	Negligible
W16	.02	–0.44	Negligible
W40	.00	1.06	Moderate

For this analysis, seven items favour students who answered in the bilingual test as indicated by the positive value for the Mantel Haenszel size. Two of the word problem DIF items favour the students answering the English-only test.

From the analysis of the SIBTEST which uses the multidimensional model, 26 items were flagged as DIF items based on the beta estimate value proposed by Roussos and Stout (1996) with statistical significance $p < .05$. From these DIF items identified, six items have been categorised as exhibiting large DIF, only two items display moderate DIF and the rest of the 18 items merely display negligible DIF. Table 3 displays the DIF items and the magnitude of Beta estimate.

For the large DIF items, four are word problem items while only two are computation items. As in the earlier methods of analysis, which identified Item W40 as moderate DIF, SIBTEST flagged W40 as having large DIF. Another item, Item C38 was identified as exhibiting negligible DIF in the previous two methods, is now flagged as large DIF. The other large DIF items, Item W35, W36, C37 and W39 were not identified as DIF in the previous two methods.

Table 3. DIF items based on the Multidimensional Model

Item	*p*-value	Beta estimate value	DIF classification	Favours
W2	0.002	0.036	Negligible	English-only
C6	0.000	0.036	Negligible	English-only
C7	0.001	0.036	Negligible	English-only
W9	0.000	0.050	Negligible	English-only
C12	0.001	0.041	Negligible	English-only
C14	0.001	0.042	Negligible	English-only
W17	0.004	0.037	Negligible	English-only
W18	0.018	0.030	Negligible	English-only
C19	0.017	0.030	Negligible	English-only
C20	0.006	0.036	Negligible	English-only
W21	0.029	0.028	Negligible	English-only
W22	0.000	0.045	Negligible	English-only
C23	0.008	0.032	Negligible	English-only
C24	0.004	0.030	Negligible	English-only
C25	0.000	0.057	Negligible	English-only
C27	0.048	0.025	Negligible	English-only
C33	0.001	0.040	Negligible	English-only
C34	0.000	0.054	Negligible	English-only
W10	0.000	0.082	Moderate	English-only
C5	0.000	0.068	Moderate	English-only
W35	0.000	–0.155	Large	Bilingual
W36	0.000	–0.180	Large	Bilingual
C37	0000	–0.169	Large	Bilingual
C38	0.000	–0.134	Large	Bilingual
W39	0.000	–0.133	Large	Bilingual
W40	0.000	–0.236	Large	Bilingual

For the two items flagged as showing moderate DIF, Item C5 and W10 were also shown to exhibit negligible DIF.

All the large DIF items seem to favour the students who answered in the bilingual test, with four of them word problem items and two computation items. In the case of moderate DIF items, both items seem to favour the students who answered in the English-only test, with one word problem and one computation item each.

To summarise the DIF analysis, by adopting IRT and CTT only one word problem item was flagged as displaying moderate DIF but using the Multidimensional Model, an additional seven items were found to display moderate or large DIF. Another notable difference is that based on IRT and CTT, the DIF item that was flagged as displaying moderate DIF, was flagged as displaying large DIF when the Multidimensional Model was used. Therefore, it

can be concluded that the Multidimensional Model flagged more DIF items as can be seen in Table 4.

Table 4. A comparison of DIF items

Mantel Haenszel Method				Multidimensional Model			
IRT	W40	Moderate DIF	Favours Bilingual test	SIBTEST	C5 W10	Moderate	All Favour English-only test
CTT	W40	Moderate DIF	Favours Bilingual test		W35 W36 C37 C38 W39 W40	Large Large Large Large Large Large	All Favour Bilingual test

As SIBTEST may be oversensitive in detecting DIF items (Stoneberg, 2004), items other than Item 40 need to be studied again to determine whether the difference detected is due to test impact or not, which does not make up the scope of this study.

Evidence of psychometric properties is further substantiated by obtaining, test statistics for both tests. The mean values of item difficulty (bilingual = 0.57, English-only = 0.58), item discrimination (bilingual = 0.57, English-only = 0.56), point-biserial (bilingual = 0.50, English-only = 0.49) and the KR 20 index (bilingual = 0.92, English-only = 0.92) for both tests indicate similar test characteristics. This observation is based on the values that are either same or with a difference of 0.01. Therefore the two tests show that they are comparable at the test level and hence, provide evidence of psychometric equivalence.

CONCLUSIONS, DISCUSSION, IMPLICATIONS AND RECOMMENDATIONS

Conclusion and Discussion

The DIF analyses between the two tests based on the CTT and IRT detected one common item which is Item 40 as exhibiting moderate DIF. The use of multidimensionality model also detected this same item as displaying large DIF and favouring students who answered in the bilingual test. In addition, another seven items were identified as displaying large DIF.

When compared to Mantel Haenszel method based on IRT and CTT, SIBTEST which is based on the multidimensional model may be oversensitive in detecting not only more DIF items as found in the studies of Boughton, Gierl and Khaliq (2000), Ercikan et al. (2004) and Stoneberg (2004) but also many more items with large DIF as discovered by Ercikan et al. (2004) and Stoneberg (2004). Therefore, the finding of this study is consistent with the findings of other studies.

However, these items must be reviewed to determine whether the DIF items are exhibiting a nuisance ability that cause unfairness (Stoneberg, 2004). As Puhan and Gierl (2006) reported, determining the cause of DIF by using experts may be challenging. In their study, three of the twelve items in their study displayed DIF for English-French speaking students where two favoured French speakers while one favoured English speakers. However, none of the four experts pinpointed translation as the cause of DIF as these items demonstrated equivalence across the two languages. Their findings show that items can be flagged statistically as DIF but expert review failed to classify them as functioning differently across the language groups**.**

In this study, it can be concluded that only one item is consistently identified as DIF item when comparing the two different version of the test. Based on the three methods of DIF analyses, CTT, IRT and the multidimensional model, only one common DIF item was flagged which made up 2.5% of the test items. As such, this suggests that the two test booklets are equivalent. Further exploration is however, recommended. This is because the presence of DIF items does not reflect test weakness since items may consistently function differently for students with special personal traits like language proficiency, as discovered in this study, which signals multidimensionality.

On the whole, it can be concluded that the additional Malay language version of the Mathematics test has not changed the traits measured. The validity of the test has not been compromised with the use of the bilingual mathematics test. Additional steps were not taken to determine whether the cause of DIF is due to bias or impact as this does not encompass the scope of this study.

Implications

The implications of these research findings are important especially for international assessment and national assessment. During international testing where Malaysia participates in TIMSS, items are translated into the Malay language. Research has suggested that teaching in the mother tongue or the first language (Bernardo & Calleja, 2005) will tremendously benefit students as it is their more proficient language. The findings of this study also reveal that the

linguistics features of the questions are reduced when mother tongue is used. However, in order to solve the Mathematics questions, an in-depth understanding of the mathematical terminologies in the language of instruction is not only necessary but a pre-requisite. As such, testing must be done in the language of instruction so that students are well familiar with the terminologies used in the classroom due to the high exposure during the Mathematics lesson.

In the national assessment, even though the bilingual test can validly assess students' mathematical achievement, linguistic simplification must be considered when constructing test items. This is to ensure that the unnecessary language load is gently removed. Among test item writers, reducing language redundancy should be an important and essential consideration when developing Mathematics test items and constructing Mathematics tests which also includes test instructions. Their accountability in doing so will ensure that language as a construct irrelevant variance that violates test validity due to its emergence as the secondary dimension that is being assessed as a part of the test construct, is removed.

Recommendations for Future Research

This research only focused on two languages which are English and Malay as these are the two dominant languages of instruction used in the Malaysian schools. The effect of other languages like Tamil and Chinese should also be investigated, as these are also languages of instruction used in the primary schools in the plural Malaysian society. The findings of this study bear great importance to test comparability. In TIMSS 2007, Malaysia continued to administer the test in the Malay language among its Form Two students which questions the test comparability of the test to a bilingual test since during this period, the language of instruction for Mathematics was in English. Future studies can look into the comparability of the two tests that can address the issue of the language of assessment for international testing among Malaysian students that can validly assess their true mathematical ability by using DIF analyses to detect items that function differently due to differences in the language.

REFERENCES

Ainan Abdul Samad (2003). *English as a tool for Science and Technology knowledge acquisition: Bilingual assessment instruments as a transitional measure*. Paper presented at the International Association of Educational Assessment, 5–10 October, Manchester, United Kingdom.

Ain Nadzimah Abdullah & Chan, S. W. (2003). Gaining linguistic capital through a bilingual language policy innovation. *South Asian Language Review*, *13*, 1–2.

Allalouf A., Hambleton, R. K. & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*, 185–198.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standard for educational and psychological testing*. Washington DC: American Psychological Association.

August, D. & McArthur, E. (1994). *Inclusion guidelines and accommodations for limited English proficiency students in the National Assessment of Education Progress*. Retrieved 2 July 2007, from http://nces.ed.gov/ pubs/p6861.pdf

Bechger, T. M. (2006). *On comparative validity*. Paper presented at The ITC 5th International Conference Psychological and Educational Tests across Language and Cultures, Brussels, Belgium, 6–8 July. Retrieved 2 April 2007, from http://www.intestcom.org/Downloads/ITC2006Brussels/ Session3.2.5TimoBechger.pdf

Bernardo, A. B. I., & Calleja, M. O. (2005). The effects of stating problems in bilingual students' first and second languages on solving mathematical word problems. *The Journal of Genetic Psychology, 166*, 117–128.

Boughton K. A., Gierl, M. J., & Khaliq, S. N. (2000). *Differential bundle functioning on Mathematics and Science achievement tests: A small step toward understanding differential performance.* Paper presented at the Annual Meeting of the Canadian Society for Studies in Education (CSSE), Edmonton, Alberta, Canada, 24–27 May.

Curriculum Development Centre (2003a). *Curriculum specifications mathematics form one*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Curriculum Development Centre (2003b). *Curriculum specifications mathematics form two*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Duncan, T. G., del Río Parent, L., Chen, W. H., Ferrara, S., Johnson, E., Oppler, S., & Shieh, Y. Y. (2005). Study of a dual-language test booklet in eighth-grade Mathematics. *Journal of Applied Measurement in Education*, *18*, 129–161.

Educational Testing Service [ETS] (2002). *ETS standards for quality and fairness*. Retrieved 2 June 2007 from http://www.efc.org/Media/About_ETS/pdf/ standards.pdf

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: sources of incomparability of English and French versions of Canada's achievement tests. *Journal of Applied Measurement, 17*, 301–321.

Ercikan, K., & Koh, K. (2005). Examining construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*, 23–35.

Elosua, P., & Lopez-Jauregui, A. (2007). Potential sources of differential item functioning in the adaptation of test. *International Journal of Testing, 7*, 39–52.

Fatimah Hashim, & Zarina Ramlan (2004). Language literacy impacting on mathematical literacy: Challenges for low proficiency ESL learners. *ETY Journal, 5*(2), 11–14.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Liberal and conservative differential item functioning using Mantel Haenszel and SIBTEST: Implications for Type I and Type II error. *Journal of Experimental Education*, *73*, 23–29.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design/ feasibility team: Report to the national assessment governing board, 1 July 1996*. Retrieved 20 May 2008, from http://www.nagb.org/pubs/appj.html

Gierl, M. J. (2004). *Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), 12–16 April, San Diego, California, U.S.A.

Gierl, M. J., & Bisanz, J. (2006). *Using test data from two Western Canadian protocol provinces to identify and evaluate factors that elicit gender differences in Mathematics*. Paper presented at the Edudata Canada Research Forum, 5 May, Vancouver, BC.

Gierl, M. J., & Khaliq, S. H. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*, 164–187.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and judgmental reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research*, *45*, 353–376.

Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H. (1996). An assessment of Stout's Index of essential unidimensionality. *Applied Psychological Measurement, 20*, 1–14.

Henrysson S. (1971). The effect of differential option weighting on multiple-choice objective tests. *Journal of Education Measurement 8, 62, 68.*

Hofstetter, C. H. (2003). Contextual and Mathematics accommodation test effects for English language learners. *Applied Measurement in Education, 16*(2), 159–188. Retrieved 1 October 2008, from http://www.informaworld.com/smpp/content ~content=a783684874~db=all

Linacre, J. M. (2008a). *Winstep*s (Version 3.67.0) [Computer Software]. Retrieved from http://www.winsteps.com

Linacre, J. M. (2008b). *A user's guide to WINSTEP MINISTEP Rasch model computer programme manual 3.67*. Retrieved from http://www.winsteps .com

Ministry of Education. (2004). *The development of education, national report of Malaysia*. Retrieved 15 December 2007, from http://www.ibe.unesco.org/ International/ICE47/english/Natreps/reports/malaysia.pdf

Messick, S. C. (1995). A validity of psychological assessment: Validation from persons' responses and performances as scientific inquiry information score meaning. *American Psychologist, 50*, 741–749.

Nelson, L. (2001a). *Lab of Educational Research Test Analysis Package* (LERTAPS) (Version 5.0) [Computer Software]. Curtin University of Technology, Perth: Assessment Systems Corporation.

Nelson, L. (2001b). *"Lelp" Lertap 5 help Interactive PDF version*. Curtin University of Technology, Perth: Assessment Systems Corporation.

Pandian, A., & Ramaiah, R. (2004). Mathematics and Science in English: *Teacher Voice*. Retrieved 11 December 2007, from http://www.melta.org .my/ET/2004/2004-50.pdf

Puhan, G., & Gierl, M. J. (2006). Evaluating the effectiveness of two-stage testing on English and French versions of a Science achievement test. *Journal of Cross Cultural Psychology, 37*, 136–151.

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality-based analysis of DIF paradigm. *Applied Psychological Measurement*, *20*, 355–371.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias? DIF as well as item bias/DIF. *Psychometrik, 58*, 159–194.

Short, D. J., & Spanos, G. (1989). *Teaching Mathematics to limited English proficient.* Received 1 October 2007 from http://www.cal.org/resources/ archives/digest/ 1989math.htm

Sireci, S. G., & Allalouf, A. (2003). Language testing. *International Journal of Testing*, *20*, 148–166.

Stoneberg, B. D. (2004). *A study of gender-based and ethnic-based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement tests applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel chi-square Test.* Retrieved 23 August 2007, from http://elearndesign, http://files.eric.ed.gov/ fulltext/ED489949.pdf

Stout, W. (2005). *Simultaneous Item Bias Test* (SIBTEST) (Version 1.7) [Computer software]. Statistical Laboratory for Educational and Psychological Measurement, Illinois: Assessment System Corporation.

Surat Pekeliling Ikhtisas Bil 11/2002. *Pelaksanaan pengajaran dan pembelajaran Sains dan Matematik dalam Bahasa Inggeris di Sekolah Kebangsaan (SK), Sekolah Jenis Kebangsaan Tamil (SJKT), Sekolah Menengah (SM) dan Tingkatan Enam mulai tahun 2003*. Kementerian Pendidikan Malaysia.

Tinkelman, S. N. (1971). Planning the objective test. In R. L. Thorndike (Ed.) *Educational measurement* (pp. 46–80). Washington DC: American Council on Education.

Yoong, S. (2005). *Teaching Science and Mathematics in English crossing linguistics & cultural borders: Problems challenges.* Retrieved 19 December 2007, from http://www.djz.edu.my/resource/maths/PDF/ speech_070708_07.pdf

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now and where it is going. *Language Assessment Quarterly*, *4*, 223–233.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel–Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.