August 2014

Assessing Mathematical Competence in Second Language: Exploring DIF Evidences from PISA Malaysian Data

Mazlina Husin
University of Wisconsin-Milwaukee

Recommended Citation

Husin, Mazlina, "Assessing Mathematical Competence in Second Language: Exploring DIF Evidences from PISA Malaysian Data" (2014). *Theses and Dissertations.* Paper 695.

ASSESSING MATHEMATICAL COMPETENCE IN SECOND LANGUAGE: EXPLORING DIF EVIDENCES FROM PISA MALAYSIAN DATA

by

Mazlina Husin

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Educational Psychology

at

The University of Wisconsin - Milwaukee

August 2014

ABSTRACT

ASSESSING MATHEMATICAL COMPETENCE IN SECOND LANGUAGE: EXPLORING DIF EVIDENCES FROM PISA MALAYSIAN DATA

by

Mazlina Husin

The University of Wisconsin – Milwaukee, 2014
Under the Supervision of Professor Bo Zhang

The year 2003 represents a significant milestone in the history of education development in Malaysia. From 2003, mathematics and science will be taught in English. This change in policy was deemed necessary to ensure that Malaysians are able to keep abreast with scientific and technological development that is mostly recorded in the English language. However, an unintended consequence of this language change was its huge impact on the national education system and the assessment of that system as well. Whenever students are not tested in their home language, one validity issue arises, which is how language, rather than the targeted knowledge, affects their performance.

The research design involves running DIF analysis for PISA 2012 mathematics assessment to verify and confirm the DIF status of the items analysed. DIF will be run using logistic regression method to check whether any mathematics items show DIF among two groups of examinees tested in their home language (Malay) and examinees tested in a second language (English). The goal is to examine whether test items functioned differently for both groups. One can investigate how the reading ability of students may affect the measurement of their performance in math. Furthermore, one can also explore whether into other important relevant variables such as socioeconomic status (SES) may explain the differential performance of students with different language backgrounds.

Mek, Chik Jusoh

Bu, Husin Abdullah

Professor Madya Dr Alias Yatim

.....Al-Fatihah

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many people I need to thank for their support while writing this thesis and during this degree. Foremost, I would like to express my sincere gratitude to my major advisor Dr. Bo Zhang for the continuous support of my study, for his patience, motivation, enthusiasm, and immerse knowledge. I would like to thank my tireless committee, who were always ready with constructive criticism and insightful comments. I would like to thank Dr. Cindy Walker, who are diligent in making sure nothing was missed. Special thanks to Dr. James Peoples, who are inspirational and always willing to help and give his best suggestions and encouragement.

A good support system is important to surviving and staying sane in grad school. I couldn't have made it through this process without my best friend Adam Malek, thank you for your everyday-without fail-pep talks. To my childhood friend Fauziah Abdullah, thank you for always believing in me more than I believed in myself. To Azrina Hadi, Alice Kones and Ramona Sedge, thank you for providing support and friendship that I needed.

I would like to thank my sisters and my brothers, who have over the years supported and guided me in my endeavors. I cannot find the words to fully express my gratitude for their unconditional love and encouragement, kept me on track all along the way particularly towards the end of this journey. Your prayer for me was what sustained me thus far. To my lovely daughter Raja Alisa Azmir, who sacrifices her innocent years, who were always there cheering me up and be by my side through ups and downs, this is for you.

Finally, I would like to thank the Malaysian Government for the financial support. Above all, I thank Allah, The Most Gracious and Merciful for His blessings. *Alhamdulillah…*

The Malaysian education system consists of four tiers: primary, lower-secondary, upper-secondary, and post-secondary throughout thirteen years of formal schooling. The year 2003 represents a significant milestone in the history of education development in Malaysia. From 2003, mathematics and science will be taught in English. This sudden emphasis on English was driven by multiple forces. This change in policy was deemed necessary to ensure that Malaysians are able to keep abreast with scientific and technological development that is mostly recorded in the English language. At the same time, this move was predicted to provide opportunities for students to use the English language and therefore increase their proficiency and be competent in the language (Education, 2002). As more and more college graduates found that jobs are limited in the domestic market, higher English proficiency would provide them with a competitive edge in international job markets.

An unintended consequence of this language change was its huge impact on the national education system and the assessment of that system as well. Whenever students are not tested in their home language, one validity issue arises, which is how language, rather than the targeted knowledge, affects their performance.

The relationship between language proficiency and mathematics achievement has been documented by a lot of researchers. According to Pearson and Champagne (2003), many teachers and curriculum specialists claim some mathematics items require students to have a high level reading ability in order to translate the reading format to correct mathematical problems. Therefore, even the students who have good mathematics background potentially will not perform well due to having a low level of reading ability. Obviously, this is the impediment to validity, if or when factors having nothing to do with the target construct (mathematics) affect examinees' scores.

Studies show that grade school students can take approximately five to seven years to acquire English language proficiency (Abedi & Gandara, 2006). Apparently, if one is not good at the language, it would be challenging for him or her to perform well on a test especially when it involves writing and reading. A test is said to be valid if it measures the construct (ability, skill, trait, or domain of knowledge) that is designed to measure the source of the examinees' scores on the test (Ferrier et al., 2011). Consequently, comparability of tests result across different language versions of these tests is a critical issue on the validity of interpretation in these assessments.

Today mathematics curricula around the world commonly include reading and communication skills, and the PISA 2012 assessment frameworks reflect this situation (OECD, 2013). For mathematics, PISA 2012 describes the theoretical assessment including mathematical literacy that assesses processes and the fundamental capabilities or competencies underlying those processes. Students should be able to solve routine and non-routine problems set in everyday contexts. Understanding the description of everyday situations for these types of problems necessarily involves reading. Furthermore, the data collected from the test items are based on "reading and interpreting" which are displayed systematically in tables, pictographs, bar graphs, and pie charts throughout the instruments.

For PISA 2012 mathematics test, the reading demands vary across items, from quite minimal, as in items requiring students to complete a computation of "naked number problem"[1], to somewhat more substantial problems, as in items requiring students to understand a phenomenon or situation and then apply their knowledge to or explain their reasoning. Often time, student's performance on mathematics items is

[1] According to Walker, Zhang, and Surber (2008), the result from "naked number problems" should not be used to label a student as proficient in mathematics, and can be considered as inaccurate because the construct of mathematics that are tested are very limited and not reflecting mathematics ability as a whole.

influenced by their level of reading ability where the format of mathematics items normally incorporates some reading comprehension that are not relative to content domain. This might contribute to item bias especially for students whose primary language is not the same with the testing language. For example, they may spend too much time trying to decode a problem thus do not have enough time or cognitive energy to comprehend (Pierce & Fontaine, 2009). In order for the achievement test scores to be valid, only their proficiency in the specific construct measured should affect students' performance. Therefore, it is important to note that unnecessary language complexity including greater emphasis on reading within subject areas should not influence students' responses to the test items.

The Malaysian case in PISA test provides a rare opportunity to study the validity of assessing mathematical competence in a second language. One can compare the performance on PISA mathematics assessment for students who were tested in home language (Malay) with those who were tested in a second language (English). The goal is to examine whether test items functioned differently for both groups. If that is not the case, one can investigate how the reading ability of students may affect the measurement of their performance in math. Furthermore, one can also explore whether into other important relevant variables such as socioeconomic status (SES) may explain the differential performance of students with different language backgrounds.

Literature Review

Linguistic Complexity of Test Items

According to Messick (1989), threats to the validity of the test score interpretation can occurs from either (a) construct-irrelevant variance (measuring something other than construct of interest) or (b) construct under-representation (incomplete measurement of the construct). Construct related evidence for validity of an assessment refers to the degree of association between the test score and what ability it is meant to describe or predict. On the other hand, Haladyna and Downing (2004), refer construct–irrelevant variance to systematic error (rather than random error) introduced into the assessment data variables unrelated to the construct being measured. Thus, we cannot make an accurate evaluation of participants' true knowledge levels.

Language testing can be considered as one of the potential sources of construct-irrelevant especially when the examinees are tested in a language that is not their native language; where it is highly likely that the proficiency to read and respond to the test items may interfere with their proficiency to demonstrate their true abilities. Schleppegrell (2004), referred linguistic complexity as "the amount of discourse (oral or written), the types of variety of grammatical structures, the organization and cohesion of ideas and, at the higher levels of language proficiency, the use of text structures in specific genres". Linguistics complexity includes such issues as the use of idioms, colloquialism, excessively long sentences or overly complicated language structures (Abedi & Lord, 2001; Kopriva et al., 2007; Wolf & Leon, 2009). Test items that consist of complicated sentences can potentially contribute to misunderstanding for some examinees. Consequently, researchers have found evidence that linguistic complexity may be hindering second language learners from having a clear understanding of the items (Abedi & Lord, 2001). Therefore, this group of students was unable to make sense of the item in order to show their ability on specific construct. Shaftel et al. (2006) in their studies on the impact of language characteristics in mathematics test items claimed that removing linguistic complexity in exam items have shown moderate increases in ELL scores compared to the original. It is suspected that linguistic complexity may leads to item bias.

As the *Standard for Educational and Psychological Testing* state, “In testing applications where the level of linguistic or reading proficiency is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid of assessment for the intended construct” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p.82). Reducing linguistic complexity on test items has been strongly encouraged by researchers as a method to increase the validity of test scores (Abedi, 2004; Abedi, 2008; Kopriva, 1999; Kopriva, 2000). These authors suggested the method of linguistic modification and linguistic simplification. Long and complicated sentence could be rephrased or replaced with a simplified sentence while still maintaining the same meaning. These methods can help to increase students’ understanding of test items by using the most common words that would be encountered in everyday conversations. In other words, this process is crucial in order to make them easier to decode as well as to avoid misunderstanding to the question being asked. Therefore, the ability level of examinees can be accurately interpreted from the test scores.

Differential Item Functioning

Fairness and equality has been a major educational theme for many years. Much emphasis has been put upon acknowledging diversity in students’ backgrounds and characteristics to ensure effective education. Test fairness is the motivation that encouraged researchers to undertake Differential Item Functioning (DIF) studies. Camilli and Shepard (1994) regarded DIF as a statistical procedure that matched examinees on the total scores to see if “comparable examinees from different groups performed the same on an individual test item" (p. 15). In other words, if an item measures the same ability in the same way across groups, regardless of the nature of the group, the same success rate should be found (O' Neill & Mc Peek, 1993). Items that give different success rates for two or more groups, at the same ability level, are said to display DIF (Holland & Wainer, 1993).

However, it is crucial to note that finding an item that displays a significant DIF is not sufficient to support the argument that the items is biased. Sometimes, an item displays DIF due to actual differences in the groups' knowledge, e.g. if one group was not taught the material therefore that group scores lower on an item, which is referred to impact. Only when the difference between the probabilities of each group passing the item is caused by construct-irrelevant factors can an item with DIF be viewed as potential biased.

In examining whether an item is biased, a lot of the literature focused on DIF detection methodology. DIF analysis can be viewed as a model-based sequential regression analysis of examinees' item responses, where item response is the dependent variable, total score is the covariate (matching variables), and the grouping variable is the independent variable. Here the 'total score' is treated as proxy to ability. This sequential regression involves a two-step modelling of item responses: (a) the *matching model*: examinees' "ability" score enters into the model first as a covariate and (b) the *full model*: the ability score, the grouping variable, and the interaction term "ability by group" enter the model. Hence, the matching model is nested within the full model. If the full model has a significant improvement in accounting for examinees' responses over and above the matching model, one can then conclude that DIF is present. In other words, an item will be flagged as having

DIF when two groups of examinees matched for their ability levels do not have the same probability of responding to an item correctly.

Swaminathan and Rogers (1990) and Zumbo (1999), state two types of DIF: uniform and non-uniform. Uniform DIF is present if one group constantly performs better than the other group across all score levels of the attribute. Uniform DIF is similar to main effect; for example, when females systematically perform better than matched males on test item. Non-uniform DIF occurs when the probability of giving a certain response to the item in the two groups is not the same for all levels of the attribute (Mellenbergh, 1982). Non-uniform DIF represents an interaction between the proficiency and performance differences across groups; for example, when high-proficiency males outperform high-proficiency females, then the pattern change to low-proficiency females outperform low-proficiency males.

Logistic regression has been widely regarded as one of the best statistical methods for evaluating DIF (Zumbo, 1999). Instead of having a normal distribution like linear regression, logistic regression uses a binomial distribution, where we are considering just one outcome variable and two states of that variable is either 0 or 1. Therefore, the probability of responding correctly to an item can be calculated for each group matched on proficiency (Zumbo, 1999). For polytomous items, where the dependent variable can be classified according to their order of magnitude, such as when the item responses with partial credit scoring, such as constructed response or short-answer test items, ordinal logistic regression model can be used. Using ordinal logistic regression has the advantage of using the same modelling strategy for binary items and DIF effect method can be extended where one has a test statistic as well as the natural corresponding measure of effect size.

Zumbo (1999) in his handbook outlined the stepwise procedure in detecting DIF using logistic regression. The first step, enters the matching or conditioning variable (total score) into the equation to account for baseline proportion of variance. In the second step, the demographically defined group (reference group and focal group) is entered. The third step, the interaction term (total score-by-group) is entered. The equation for logistic regression is:

$$Y = \beta_0 + \beta_1(total) + \beta_2(group) + \beta_3(total \times group)$$

where Y is a natural log of the odds ratio. That is the equation:

$$ln\left[\frac{P_i}{1 - P_i}\right] = \beta_0 + \beta_1(total) + \beta_2(group) + \beta_3(total \times group)$$

where P_i is the probability of responding to item *i* correctly, $1 - P_i$ refers to the probability of responding to item *i* incorrectly, $\beta 1(total)$ is the regression coefficient for the matching or conditioning variable (i.e. total score), $\beta 2(group)$, is the regression coefficient for group membership (dummy coded as 0 = reference group, 1 = focal group), and $\beta 3(total \times group)$ is the regression coefficient for the interaction between group and matching variable.

The test of the DIF significance can be calculated by taking the chi-square for the total score and deducting it from the chi-square of the interaction and using the chi-square table to compare the results with two degrees of freedom. A significant main effect for group membership and the interaction between group membership (reference group and focal group) and ability level (total score) in the regression indicates that ability level alone does not predict the successful of answering the item correctly. A significant interaction means that the DIF is non-uniform and that the slopes differ for the groups where their regression lines may cross; that suggest the item favors one group either at the higher or lower end of the ability.

According to the effect size classification initially suggested by Zumbo (1999), moderate DIF will yield an R^2 between 0.13 and 0.26, while for large DIF the R^2 should exceed 0.26. However, Jodoin and Gierl (2001) were concerned that Type 1 errors might increase as sample size increase. They proposed new guidelines for logistic regression and used by Educational Testing Service, items can be classified as displaying negligible or A-level DIF ($R^2 < 0.035$), moderate or B-level DIF (null hypothesis is rejected and $0.035 \leq R^2 < 0.07$), or large or C-level DIF (null hypothesis is rejected and $R^2 \geq 0.07$).

A growing number of DIF studies have researched situations in which comparable test-takers from diverse ethnic, racial, cultural, or linguistic backgrounds have had different probabilities of success on a given item on standardized achievement or proficiency (Geirl & Khaliq, 2000; Kim & Jang, 2009; Klieme & Baumert, 2001), and DIF results have been used to enhance the quality of the studies. Furthermore, choosing more than one DIF method and considering item as DIF as long as they are simultaneously detected across all the statistical analyses used, would reduce the error rate to a certain extent (Camilli & Shepard, 1994).

DIF detection methodology has been previously used in several studies to evaluate the effect of testing students in a secondary language. For example, Yildirim and Berberoglu (2009) used DIF analyses to evaluate content-wise evaluation between test takers from the United State and Turkey. Using both substantial and statistical analyses, they found that three sources of errors that cause DIF in PISA. These sources are as follows: mathematics literacy items, translation errors, and use of quantitative words. Based on another large-scale assessment, Arim and Ercikan (2005) have reported that 23% of items in Trends in International Mathematics and

Science Study (TIMSS) displayed DIF when English and Turkish speaking examinees were compared.

Similarly, one of the most well-known studies on test language is by Chen and Henning (1985) using Rasch model and regression procedure. They investigated the extent to which items on the English as a Second Language Placement Examination (ESLPE) functioned differently for students whose native language was Chinese and Spanish. Result showed that out of 150 items, four vocabulary items were flagged for DIF against Chinese students and those words were more familiar to Spanish students. Familiarity provided them better chance to make sense of the question asked.

Schmitt (1988) used DIF analyses to identify items on a college admissions test that functioned differently between two groups of students, Euro-American and Hispanic-American. Items flagged for DIF were found to have terms that differed with respect to familiarity across these two groups. Gierl, Rogers, and Klinger (1999), have reported that 52% of items in a Canadian achievement test displayed DIF across English and French speaking examinees.

Most previous studies have analysed DIF in multiple-choice questions. One interesting study on constructed response test by Lee et al. (2005) investigated writing prompts in the test of English as a Foreign Language (TOEFL) between European and East Asian language group. Using a logistic regression method, they found that prompts flagged for DIF had very small effect size and conclude that the writing prompts were not biased against both groups. Another DIF study comparing Caucasians and minority ethnic groups have found that open-ended (constructed response) test items favour minorities while multiple choice items favour Caucasians (Taylor & Lee, 2011).

Every achievement test must carefully include appropriate words and examples to avoid miscomprehension, particularly to second language learner. Kopriva (2000), suggested using high frequency words to reduce the cognitive reading load so the students can concentrate on the task and demonstrate their skill in the content area. Apart from choice of words, item length can also be associated with language complexity in assessing second language learners. Shaftel et al. (2006) found that longer test items are more difficult than shortest test items on a State Mathematics Assessment at three different grade levels for second language learners. Abedi and Lord (2001), in their study used DIF detection procedures to determine whether simplifying the English language on math test items led to performance differences for non-native learners.

Analysis of DIF has contributed to important interpretations of how language proficiency in the language of the test, affects students' test performance. Therefore, item developers must investigate and pay attention to the sources of error (i.e., linguistic complexity) during the item building process because it would be economically and technically worthwhile if it were possible to minimize construct-irrelevant variance and detect items with potential DIF before the test is administered. Particularly, when score interpretations are made with respect to entire country.

In addition, there were also studies that included other extraneous variables (i.e. cultural or background variables) in addition to the ability being measured to identify DIF more accurately. For example, Clauser et al. (1996) were able to confirm that extra matching on an educational background variable could improve the precision of detection of DIF items in the National Board of Medical Examiners' Part III examination. However, researchers frequently focus on translation and content area; only few actually go beyond these factors to investigate other cultural sources of

DIF. Therefore, it is crucial to include other extraneous variables such as students' reading ability and socioeconomic status in the investigation and see if they are likely to affect score comparability across the groups of interest.

Socioeconomic Status

Socioeconomic status (SES) is typically defined by family income, level of poverty in the child's neighborhood, educational attainment by parents, and occupation of the heads of households (Clements & Sarama, 2008). Colemam (1966), in his study on Equality of Educational Opportunity claimed that the influence of experiences that a student is exposed to may depend to a large degree on family background which are greater than anything that goes on within schools.

Many studies found that socioeconomic status, the level of family income; low SES or high SES, has been seen as a strong predictor of student academic achievement across the nation (Coleman, 1966; White, 1982; and Klingele & Warrick, 1990) and is associated with large differences in performance in most countries and economies that participate in PISA. Socio-economically advantaged students and schools tend to outscore their disadvantaged peers by larger margins than between any other two groups of students (OECD, 2013).

Research that compares high SES students to low SES students has revealed poorer educational outcomes can occur due to: lack of parental involvement, lower parental education level, less school resources, lack of the availability of advanced placement courses in high school and overall differences in content covered in class lessons (Schmidt, Cogan, & McKnight, 2011). Researchers have also examined the effect of SES on mathematics achievement. Most important, early influences of SES appear to be greatest on verbal aspects of mathematics (Jordon et al., 2007). On average, children from disadvantaged low-income families perform substantially worse in mathematics than their counterparts from higher income families (as reviewed by the National Mathematics Advisory Panel, 2008).

Conversely, many socio-economically disadvantaged students succeed at school, and many achieve at high level on the PISA assessment. In fact, many countries and economies that have seen improvements in their mean performance on PISA have also managed to weaken the link between socio-economic status and performance (McConney & Perry, 2010). Although having low SES does not guarantee a negative effect on academic performance, they are considered as a dominant trend that can be associated to unfavorable educational outcomes. Thus, having a complete dataset that consist of demographic information such as SES and other background characteristics are extremely important for researchers in determining effective and valid testing for all students (Kopriva, Wiley, and Emick, 2007). It is crucial for all educators and item developers to understand, so that all the students can achieve to their academic potential.

The diversity among students should be taken into consideration when interpreting each student's proficiency as well as to run comparisons within or between groups of interest. However, many studies suffer from not having access to demographic information that would improve on their results. Therefore, matching scores in a DIF analysis for SES could be an important component in comparing the results of academic performance and reveal other factors that could cause score variance, especially for the studies that conclude certain learning aspect or content area are the cause of DIF.

Research Questions

1. Is DIF present for math items among group of students who were tested in their home language or in a second language?
2. Does reading level play a significant role in DIF, beyond group membership?
3. Does SES play a significant role in DIF, beyond reading level and group membership?

Methods

Research Design

To answer the above research questions, the research design involves running DIF analysis for PISA 2012 mathematics assessment to verify and confirm the DIF status of the items analysed. DIF will be run using logistic regression method to check whether any mathematics items show DIF among two groups of examinees tested in their home language (Malay) and examinees tested in a second language (English). For items that show DIF, a second run will be conducted while controlling the reading ability of students. The data will be further analysed to detect if SES has a significant affect in DIF beyond reading ability.

Zumbo (1999) suggested that sample sizes be 200 or larger when using logistic regression to evaluate items for DIF. For purposes of this analysis, combining three available data from OECD PISA 2012 (students, parents, and school questionnaire), our sample consists of 5197 observation that met this criterion across all types of schools in Malaysia (i.e., Fully Residential School, National Secondary School, Religious School, Technical School and others).

PISA Data

The analyses in this study use data from the 2012 wave of the Program for International Student Assessment, referred to as PISA. PISA 2012 is the program's 5th survey, collected by the Organization for Economic Co-operation and Development (OECD). First administered in 2000, PISA is a survey developed jointly by participating countries all around the world. This program investigates and compares the performance of schools and education systems in all thirty four (34) OECD member countries and thirty one (31) partner countries by assessing the competencies of 15-year-olds in three main subjects: Mathematics, Reading, and Science. PISA attempts to measure students' capacities to apply knowledge and skills, using assessment tasks involving multistep reasoning and real-world situations, as opposed to mastery of a particular curriculum.

PISA is a complex survey data. Data were collected from nationally representative samples of students and their principals in a two-stage, stratified, cluster design. Schools that participate in the survey have been chosen first, being therefore considered as the primary units. Schools were sampled systematically with probabilities proportional to size, the measure of size being a function of the estimated number of the eligible (15-year-old) students enrolled in the school. In the second stage, a random sample of students from the target population was drawn from every selected school. The sample is representative of the target population.

Students were given an instrument of standardized achievement test to assess their mathematical, reading, and science literacy and the questionnaires that asked a number of questions about themselves, their attitudes and approaches to learning, personal characteristics including socioeconomic status and language spoken in the home and also regarding their schools. The administrators of the schools or the principals also answered questionnaires to provide contextual information describing the students and their families, their schools characteristics such as facilities and resources, instructional process and climate. Approximately 510,000 students between the ages of 15 years 3 months and 16 years 2 months participated in the assessment were selected to take a standardized test (OECD, 2012) representing about 28 million 15-year-olds globally.

The assessments are held every three years, and each round places a special focus on one of the key subjects. For PISA 2012, the major subject was mathematics literacy, defined as "an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain, and predict phenomena. It assists individuals to recognize the role that mathematics plays in the world and to make the well-founded judgements and decisions needed by constructive, engaged, and reflective citizens" (OECD 2013, p.25).

According to the PISA framework, individual achievement in mathematics literacy is measured by a scaled score adjusted for reliability, difficulty and guessing, using Item Response Theory statistical procedures (Hambleton, Rogers, & Swaminathan, 1991). The PISA mathematical literacy proficiency scale comprises six levels of progressions (Level 1 - 6). At the highest level (i.e. Level 6), students are capable of advanced mathematical thinking and reasoning, and can apply this insight and understanding, along with a mastery of symbolic and formal mathematical operations and relationships so as to develop new approaches and strategies for attacking novel situations. Furthermore, students can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situations (OECD, 2013).

On the other hand, bottom performing students (i.e. Level 1) can only answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. Students at this level are able to identify information and to carry out routine procedures according to direct instructions in explicit situations, and can perform actions that are obvious and follow immediately from the given stimuli (OECD, 2013).

This scale is defined in relation to three dimensions for purposes of test development: (1) content categories; (2) context categories; and (3) mathematical processes. Instead of commonly used curricular components such as numbers, algebra and geometry, overarching ideas reflecting orientation toward real life situations are used to define the PISA test contents. Students took a paper-based test that lasted approximately two (2) hours. The tests provides problems in a variety of item formats, a mixture of multiple-choice questions and open-ended questions; each had four or five options that were organized in groups based on a passage settling out a real-life situation.

PISA employed matrix sampling procedures where students responded to achievement items from thirteen (13) different booklets[2] (students took different combinations of the different tests). PISA provides five plausible values per scale or subscale. If an analysis is to be undertaken with one of these five cognitive scales then the analysis should be undertaken five times, once with each of the five relevant plausible values variables. The results of these five analyses are averaged and then significance tests that adjust for variation between the five sets of results are computed.

[2] To reach satisfactory coverage, many items need to be developed and included in the final test. At the same time, it is unreasonably to assess a sampled student with the whole instrument; therefore PISA implements a rotated test design (see OECD, 2001 initial report).

Item Classification

Three dimensions classifying the item characteristics which were defined in the PISA framework (OECD, 2012), were the main focus for examining the patterns of DIF in this study. The detail categories are as follows: (1) *Content Categories*: Space and shape, change and relationships, quantity, and uncertainty and data. (2) *Context Categories*: Personal, occupational, societal, and scientific. (3) *Mathematical Processes*: Formulating situations mathematically, employing mathematical concepts, facts, procedures, and reasoning, and interpreting, applying and evaluating mathematical outcomes.

The current PISA test consists of two types of cognitive items: (1) Multiple choice: simple multiple choice and complex multiple choice; that is a series of true/false or yes/no choices, one answer to be chosen for each element in the series; and (2) Construct response, most of items require markers. The data of the items were recoded as dichotomous (0 and 1) and partial credit (0, 1, and 2).

Dependent and Independent Variables of the Study

For the purpose of this study, the dependent variable was the score they received on the each item (both multiple choice and construct response format) and the matching or independent variables were the PISA 2012 five plausible values for mathematics, five plausible values for reading, students' socioeconomic status (SES), and the group membership which was determined by language of the testing instrument that the examinees responded, either home language (Malay) or second language (English). Reading scores and SES will also be used as independent variables in DIF analyses.

Mathematics Score (PVMATH)

The primary predictor of interest is mathematics performance, a mathematic scale score treated as a proxy for ability, measured at the individual level and estimated with five plausible values (PV1MATH…PV5MATH). In mathematics, PISA measures students' ability to activate their knowledge and skills to solve problems found in real-life situations. It centres around three major domains of assessment: mathematics content categories, mathematics contexts, and mathematical processes.

Similar to mathematics, reading literacy[3] scale score is measured at individual level and also treated as a proxy for ability, measured at the individual level and estimated with five values (PV1READ…. PV5READ). Reading literacy includes a wide range of cognitive competencies, from the basic decoding to knowledge of words, grammar and larger linguistic and textual structures and features, to knowledge about the world. Examinees need to exhibit their understanding, using reflecting on and engaging with written texts, in order to achieve one's goals, develop one's knowledge and potential, and participated in society (OECD, 2013). The PISA reading literacy assessment is built on three major task characteristics to ensure a broad coverage of the domain: situation, text and aspects.

The simplest way to describe plausible values[4] is to say that plausible values are some kind of student ability estimates. It is very important to be aware that

[3] According to PISA 2012 framework, reading literary assessment domains are: (1) situation – refers to range of broad contexts or purposes for which reading takes place (2) text – the range of material that is read (3) aspect – refers to the cognitive approach that determines how readers engage with a text.

[4] Wu (2005) explained that in large-scale assessment programs such as the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and PISA, achievement data sets provided "plausible values". Those values can be used to: (1) address concerns with bias in the estimation of certain population parameters when point estimates of latent achievement are used to estimate those population parameters, (2) allow secondary data analysts to employ "standard" techniques and tools to analyze achievement data that contain measurement error, and (3) facilitate the computation of standard errors of estimates when the sample design is complex

plausible values are not test scores and should not be treated as one. Plausible values are random numbers that are drawn from the distribution of scores that could be reasonably assigned to each examinee (Monsuer and Adams, 2009). Those values were developed to obtain consistent estimates of population characteristics[5] where examinees are administered too few items to allow precise estimates of their ability. Using IRT scaling to estimate scores, as in PISA and many large-scale assessment, each students receives a subset of the total set of items. These procedures enable test designers to include a substantially larger number of items than would be feasible for examinees to complete.

Group (GROUP)

PISA 2012 provides dataset with the full set of responses from questionnaires to individual students, parents, teachers, cognitive item response and scored cognitive item response. In the instrument of academic achievement tests, PISA asked all students to identify the language of the test used to answer the test questions.

For the Malaysian survey, students will be divided into two groups (reference group and focal group) based on the test languages selection they made, Malay (204) and English (313). Table 1 presents the number of students tested in Malaysia, divided into two groups based on the language of the test. 75.1% of the examinees answered the questions in home language and 24.8% answered in second language.

[5] The assignment of valid and reliable scores to individuals is not a purpose of PISA, but to describe populations.

TABLE 1
Sample Size for Students Groups

Group *(Language of the Test)*	*Total # of Examinees*	
	n	%
Home Language (204)	3905	75.1%
Second Language (313)	1290	24.8%
Missing	2	0.1%
Total	5197	100%

Note. Language of the test: 204 = Malay; 313 = English

Individual/Student Socioeconomic Status (SES)

Student SES[6] in PISA is termed "educational, social, and cultural status" (ESCS). ESCS variable was derived from student responses to questions about the following three indices: highest occupational status of parents (HISEI[7]), highest educational level of parents in years of education according to (ISCED[8]), and home possessions (HOMEPOS[9]). The final values on the PISA index of ESCS for 2012 have an OECD mean of 0 and a standard deviation of one. ESCS is thus a comprehensive and detailed measure of individual student SES. For Malaysian data, values on the index range from -4.11 to 1.86 with higher values representing higher socioeconomic status.

[6] SES variable is derived from item about "annual household income" in the background questionnaire. However, out of 65 countries that took part in the PISA 2012 survey, only 11 countries provide the income data, 3 were from Asian countries (Korea, China-Macao, and China-Hong Kong). For Malaysian data, the variable was coded as 'a' which means that the category does not apply to the country concerned, data therefore missing.

[7] HISEI: the index is designed to optimize equivalence in occupations across countries. Those occupations are: elementary occupations, semi-skilled blue-collar occupations, semi-skilled white-collar occupations and skilled occupations.

[8] ISCED: the index is derived from parents' level of educations: tertiary education, secondary educations as their highest level of education, attained other post-secondary qualifications.

[9] HOMEPOS: the index comprises of all items on the indices of WEALTH, CULTPOSS and HEADRES such as works of classical literature, works of art (e.g. paintings), as well as books in the home recorded into a four-level categorical variable (0-10 books, 11- 25 or 26-100 books, 101-200 or 201-500 books, more than 500 books). Students are asked how many bedroom, computers, book and original artworks are in their home, and how often they visit museum, art galleries, and concert halls.

DIF Detection

In this study, we applied a two-step modelling procedure for each item by the following three steps.

Step 1: *Matching Model*: only ‘math’ entered the model as a covariate first.

$$logit = \beta_0 + \beta_1(math)$$

(1 degree of freedom)

Full model: ‘math’ entered the model as a covariate first, and then ‘group’ and the interaction entered the model.

$$logit = \beta_0 + \beta_1(math) + \beta_2(group) + \beta_3(math * group)$$

(3 degrees of freedom)

Step 2: *Matching model*: two matching variables, ‘math’ and ‘group’ and their interaction term went into the model as covariate first.

$$logit = \beta_0 + \beta_1(math) + \beta_2(group) + \beta_3(math * group)$$

(3 degrees of freedom)

Full model: in addition to the two terms in the matching model, ‘read’ and the interaction of ‘math’ and ‘group’ enter the model.

$$logit = \beta_0 + \beta_1(math) + \beta_2(group) + \beta_3(reading) + \beta_4(math * group)$$

(4 degrees of freedom)

Step 3: *Matching model*: three matching variables, ‘read’, ‘math’ and ‘group’ and their interaction term went into the model as covariate first.

$$logit = \beta_0 + \beta_1(math) + \beta_2(\text{group}) + \beta_3(reading) + \beta_4(math * \text{group})$$

(4 degrees of freedom)

Full model: in addition to the three terms in the matching model, ‘SES’ and the interaction of ‘math’ and ‘group’ enter the model.

$$logit = \beta_0 + \beta_1(math) + \beta_2(\text{group}) + \beta_3(reading) + \beta_4(SES) + \beta_5(math * group)$$

(5 degrees of freedom)

For Step 1, an item would be flagged as DIF by statistically testing the difference in Chi-square values between the matching and full models at $\alpha \leq 0.05$ level with 1*df*. For Step 2, the Chi-square difference between the matching and the full models was tested for significance at $\alpha \leq 0.05$ level with 1df. If READING is a source of DIF, we expect that the number of DIF items detected at Step 1 would decrease. For Step 3, the Chi-square difference between the matching and the full models was tested for significance at $\alpha \leq 0.05$ level with 1df. If SES is a source of DIF, we expect that the number of DIF items detected at Step 2 would further decrease.

However, in this study we were only interested in whether an item was detected as DIF at the two models in the three Steps and the direction of the values for detecting group favouring of DIF rather than the form (i.e. uniform DIF or non-uniform DIF).

Results

The descriptive statistics for five plausible values for mathematics are presented in Table 2. The largest mean difference between the two groups of interest was PVM_4 (62.34) and the smallest mean difference was PVM_1 (60.96), which was two-third of standard deviation. Similarly, the descriptive statistics for plausible values for reading on Table 3 shows that the largest mean difference between the two groups of interest was PVR_4 (24.83) and the smallest mean difference was PVR_2 (22.46), which was one-fourth of standard deviation. For this data it is interesting to note that the examinees who were in second language group (English) had higher average plausible values for both mathematics and reading than the home language group (Malay).

TABLE 2
Descriptive Statistics by Group and 5 Plausible Values for Mathematics

	Group (Language of the Test)		
Plausible values	*Home Language (SD)*	*Second Language (SD)*	*Mean Difference*
PVM_1	407.68 (71.83)	468.64 (88.67)	60.96
PVM_2	406.81 (72.10)	468.71 (87.30)	61.90
PVM_3	406.49 (71.70)	468.62 (87.18)	62.13
PVM_4	405.99 (71.68)	468.33 (87.83)	62.34
PVM_5	406.11 (72.28)	468.41 (86.30)	62.30
Average	406.61	468.54	

Note. Language of the test: 204 = Malay; 313 = English; PVM = Plausible values for Mathematics

TABLE 3
Descriptive Statistics by Group and 5 Plausible Values for Reading

	Group (Language of the Test)		
Plausible values	*Home Language (SD)*	*Second Language (SD)*	*Mean Difference*
PVR_1	394.57 (78.16)	418.68 (93.80)	24.11
PVR_2	395.12 (78.33)	417.58 (91.72)	22.46
PVR_3	394.39 (78.60)	417.96 (91.94)	23.57
PVR_4	393.30 (78.20)	418.13 (93.26)	24.83
PVR_5	394.47 (78.96)	417.67 (92.19)	23.20
Average	394.37	418.01	

Note. Language of the Test: 204 = Malay; 313 = English; PVR = Plausible values for Reading

As we can see from Table 4, the number of DIF items were higher in construct response format, 20 (62.5%) in Step 1, 9 (60%) in Step 2 and 4 (57.1%) in Step 3. This may indicate that, there was a possibility that DIF was due to differences in the item format between the two groups.

TABLE 4
Number of DIF Items by Item Format

	Format of the Items		
Steps	*MCQ (%)*	*Construct (%)*	*Total*
Step 1 (PVM)	12 (37.5%)	20 (62.5%)	32
Step 2 (PVM + PVR)	6 (40.0%)	9 (60.0%)	15
Step 3 (PVM + PVR + SES)	3 (42.9%)	4 (57.1%)	7

Note. DIF = Differential Item Functioning; PVM = Plausible values for mathematics; PVR = Plausible values for reading; SES (socioeconomic status) is the index for ESCS = educational, social, and cultural status; MCQ = Multiple choice questions.

Three main research questions are used to guide the reporting of data analyses. The questions were based on the issues concerning mathematics performance differences between home language and second language group due to reading ability, and SES. We hypothesized that reading ability was the source of DIF in the PISA 2012 mathematics score comparison between examinees who took the test in home language and second language. This hypothesis was tested by looking at the number of DIF items when the additional matching variable is included in the model. Table 5 shows the result in terms of number and percentage of items that display DIF. As we can see on Step 2, when extra matching variable PVRead is included in the models in addition to PVMath, we expected that some items flagged as DIF on Step 1 would no longer show DIF or the total number of items would decrease. The results were consistent with our hypothesis. From 32 items that display DIF on Step 1, the number decreased to 15 items, decreased by 53.1%.

We further investigate SES variable to see if it has a significant role in DIF, beyond the group membership and reading level. As we can see on Step 3, when extra matching variable SES is included in the models in addition to group memberships, PVMath, and PVRead, we expected that some items flagged as DIF on Step 2 would no longer show DIF or the total number of items would decrease. Similarly, the results were also consistent with our hypothesis. From 15 items that display DIF on Step 2, the number decreased to 7 items, decreased by 53.3%.

TABLE 5
Summary of Differential Item Functioning (DIF) Analysis by Steps

Steps	*# of item showing DIF*	*% of DIF Reduction*
Step 1 (PVM)	32	
Step 2 (PVM and PVR)	15	17/32 (53.1%)
Step 3 (PVM, PVR and SES)	7	8/15(53.3%)

Note. DIF = Differential Item Functioning; Language of the Test: 204 = Malay; 313 = English; PVM = Plausible values for mathematics; PVR = Plausible values for reading; SES (socioeconomic status) is the index for ESCS = educational, social, and cultural status.

In addition to comparisons in the reduction of DIF status between the three steps, we looked further into the patterns of three domains of mathematics literacy (i.e. content categories, context categories and mathematical processes categories). As we can see from Table 6, in Step 2, PVRead reduced the DIF status of all subscales in all the three domains. For content domain, PVRead reduced the DIF status for five out of nine (55.6%) items from “Change and relationships” subscale and “Uncertainty and data” subscale. For context domain, PVRead reduced the DIF status for seven out of twelve (58.3%) items from “Scientific” subscale. For mathematical processes domain, PVRead reduced the DIF status for nine out of eighteen (50%) items from “Employing concepts” subscale and five out seven items from “Formulating situations” subscale. This may indicate that reading ability did have an effect on whether an item was flagged as DIF in those specific subscales.

TABLE 6
Number of DIF item at Different Steps with Different Domains

Major Domain of Mathematic Assessment	*Steps*			*DIF Reduction (%)*	
	PVM	*PVM+PVR*	*PVM+PVR+SES*	*Step 1 to 2*	*Step 2 to 3*
Content categories:					
Space & shape	7	3	2	4 (57.1%)	1 (33.3%)
Change & relationships	9	4	1	5 (55.6%)	3 (75.0%)
Quantity	7	4	1	3 (42.9%)	3 (75.0%)
Uncertainty & data	9	4	3	5 (55.6%)	1 (25.0%)
Total number of items:	32	15	7		
Context categories:					
Personal	6	3	1	3 (50.0%)	2 (66.7%)
Occupational	2	1	0	1 (50.0%)	1 (100%)
Societal	12	6	5	6 (50.0%)	1 (16.7%)
Scientific	12	5	1	7 (58.3%)	4 (80.0%)
Total number of items:	32	15	7		
Mathematical processes:					
Formulating situations	7	2	0	5 (71.4%)	2 (100%)
Employing concepts	18	9	5	9 (50.0%)	4 (44.4%)
Interpreting outcomes	7	4	2	3 (42.9%)	2 (50.0%)
Total number of items:	32	15	7		

Note. DIF = Differential Item Functioning; Language of the Test: 204 = Malay; 313 = English; PVM = 5 Plausible values for mathematics; PVR = 5 Plausible values for reading; SES (socioeconomic status) is the index for ESCS = educational, social, and cultural status.

Having additional matching variable SES in Step 3, similar to the analysis on Step 2, we found that it also reduced the DIF status of all subscales in all three major domains of mathematical assessment. SES reduced the DIF status for four out of nine (44.4%) items from “Employing concepts” subscale and four out of five (80%) items from “Scientific” subscale. This may indicate that, SES was largely related to whether an item was flagged as DIF between the two groups and can be seen as a source of DIF in those domains.

Since we do not have the access to all items, we could only analyze and reported the items that appeared in the PISA 2012 Released Items. Example of one DIF item that no longer show DIF on Step 2 (after controlling for PVMath and PVRead) was examined to determine what may have caused the DIF between the two groups. This item fell in the subscales of space and shape, scientific and employing mathematical concepts. Item 80 asked: “*What is the size in degrees of the angle formed by two door wings?*” The result may indicate that language complexity has an effect on whether an item was flagged as DIF in that domain. To confirm, a score analysis was conducted to see the pattern of response between the two groups. We found that 44.59% of examinees from English group answered the question correctly compared to only 32.55% from Malay group (see Figure 1).

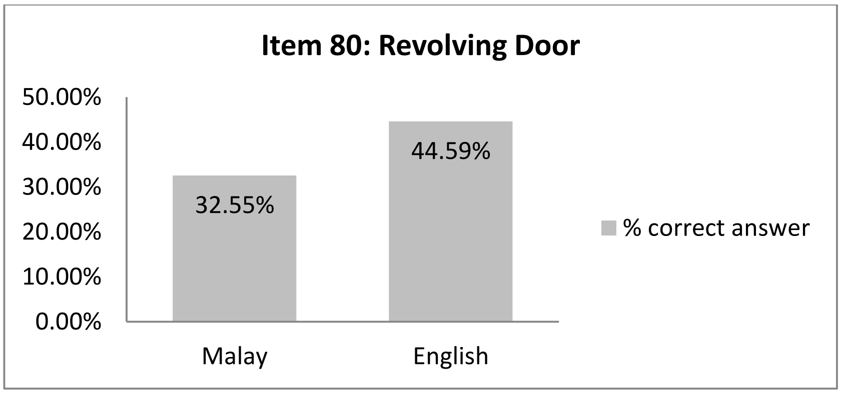

FIGURE 1:
Score Analysis for Item 80

As we can see from parameter estimate on Table 7, the coefficient for group variable is 0.148 which indicates that the item is favouring the second language (English) group. It is interesting to note that the item stem included the word “wings” which could be confusing for some examinees who answered the question in home language (Malay) because the word “*daun*” is more relevant to parts of plants rather than the “door”. This finding indicated that there was a possibility that DIF was due to the unfamiliarity of words in the question. However, a more detailed analysis of word familiarity would be needed to support that hypothesis.

TABLE 7
Parameter Estimate for Group on Step 2 by DIF Items

Item	*Restricted Model*		*Full Model*	
	Group	*(S.E)*	*Group*	*(S.E)*
4	0.400	(0.143)	0.362	(0.143)
9	-0.058	(0.151)	-0.033	(0.151)
16	-0.173	(0.115)	-0.164	(0.116)
33	-0.331	(0.115)	-0.351	(0.116)
36	-0.26	(0.113)	-0.248	(0.114)
44	-0.696	(0.189)	-0.657	(0.189)
45	0.275	(0.128)	0.338	(0.127)
49	0.604	(0.142)	0.629	(0.143)
50	0.034	(0.106)	0.079	(0.106)
53	0.322	(0.124)	0.368	(0.124)
55	-0.493	(0.138)	-0.523	(0.139)
56	-0.279	(0.144)	-0.323	(0.145)
59	-0.126	(0.132)	-0.151	(0.133)
73	0.244	(0.121)	0.284	(0.121)
74	-0.171	(0.137)	-0.124	(0.136)
80[10]	0.147	(0.131)	0.148	(0.132)

On the other hand, Item 33 favors the home language (Malay) group with the coefficient of -0.351. The item asked question about Number Check that fell in the subscales of quantity, scientific and employing mathematical concepts. From the score analysis, we found that 56.8% of examinees from Malay group answered the question correctly compared to 43.2% from English group (see Figure 2).

[10] Item 80 does not belong to the 15 items that show DIF on Step 2. It is included in Table 7 to show the coefficient value.

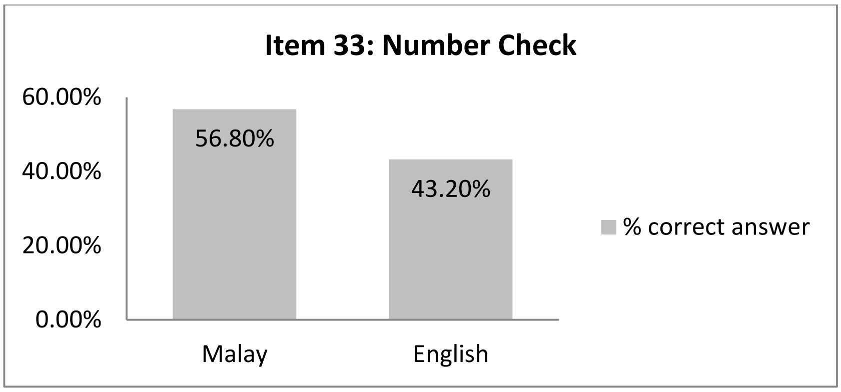

FIGURE 2
Score Analysis for Item 33

Similarly, we further investigated DIF items that no longer show DIF on Step 3 to determine if SES has a significance effect on what may have caused the DIF beyond the reading level and group membership. Item 50 asked: "*How many CDs did the band The Metalfolkies sell in April?*" The result may indicate that unfamiliarity of the word "CDs" has an effect on whether an item was flagged as DIF. From the score analysis, we found that 86.73% of examinees from English group answered the question correctly compared to 85.8% from Malay group (see Figure 3). The coefficient for group variable for these items are 0.079(see Table 7) which indicates that the item is favouring the second language (English) group.

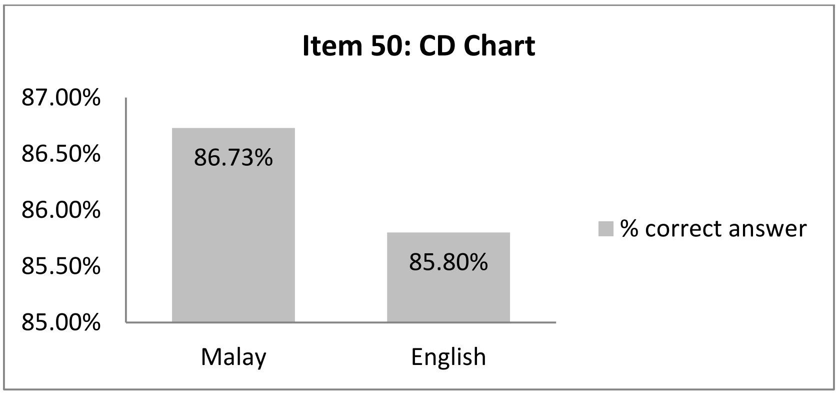

FIGURE 3
Score Analysis for Item 50

TABLE 8
Descriptive Statistics by Group and SES

Group *(Language of the Test)*	*Mean (SD)*
Home Language (204)	-.90 (.95)
Second Language (313)	-.21 (.91)
Total	-.72 (.99)

Note. Language of the Test: 204 = Malay; 313 = English; SES (socioeconomic status) is the index for ESCS = educational, social, and cultural status.

Based on descriptive statistics by group and SES (see Table 8), it was found that on average, the SES index for English group (-.21) is higher than the Malay group (-.90). This finding indicated that there was a possibility that DIF was due to the examinees level of SES. For instance, examinees from lower SES level may not be familiar or may not expose to "CDs". However, it will be helpful to evaluate all items that show DIF due to SES in order to support that hypothesis.

From the analysis on Table 7, the result shows that out of fifteen items that show DIF, nine items were favouring the home language (Malay) as compared to only six items were favouring the second language (English) group. These results were consistent with our hypothesis that when examinees are tested in the language they are not familiar with, the proficiency to read and respond to the test questions may interfere with their proficiency to exhibit their knowledge, skills and abilities.

Discussions

Differential item functioning by subsamples, in particular language of the test DIF, is unavoidable in large-scale tests such as in PISA, TIMSS and many others. According to Holland and Wainer (1993) in Sireci (1997), issue in assessing students who operate in different languages are among the most difficult problems facing contemporary psychometricians. For instance, SAT test that is required as part of

College admissions process, the results of the exam involve high-stakes decision about the test-takers with different language backgrounds. As the use of the score is directly related to their acceptance to a university, test users must be particularly cautious in ensuring that test scores are interpreted accurately and to avoid unnecessary hardship for test-takers (Gennaro, 2006).

The problem will be more challenging in standardize achievement test that involve more than one language. There is a greater need for the tests to reflect the accuracy especially in cases involving a diverse population, including students who answer the test item in their second language. This is because, when students from different language background respond to test questions in different languages, it is difficult to establish construct equivalence – the trivial factor that contributes to test equality.

In this study, we investigated whether students' familiarity with the language in which a test was administered affected their performance on mathematics test items after they were matched on overall test performance. The results showed that the overall test performance may be explained partly by language factors, particularly on the three domains as well as the item formats and background variable such as student's SES.

These findings suggest that issue related to language factors in assessing mathematics in second language learner is necessary in international test especially when the result is used to rank the participating countries. Additionally, their language background variables should always be considered, and efforts should be made to reduce confounding effects to ensure accurate assessment outcome.

We also found several interesting methodological strategies that can be used to evaluate DIF on large-scale international assessment. Firstly, students' plausible values can be used as the matching criterion across groups when test results are computed using that methodology. Working with five plausible values may seem overwhelming or cumbersome and some researchers analysed the data incorrectly and tried to resort to shortcuts like using just one of them to simplify the calculation of means and variances or other analyses that lead to biased results. For instance, Hauger and Sireci (2008) only used one out of five plausible values provided for each students in their study on DIF across examinees from three countries.

As mentioned, plausible values are not a test scores, those are random numbers that are drawn from the distribution of scores that contain random error variance components and are not accurate as scores for individuals. Carstens and Hastedt (2010), in their study on the effect of not using plausible values the correct way using TIMSS 2007 grade 8 mathematics data, shows that inappropriate use of the plausible values or alternative scoring methods can lead to the risk of producing biased estimates, underestimates of standard errors, or inferences that are not supported by the data.

Although plausible values are a convenient criterion, one other area of potential future research is to incorporate different matching variable that are available in demographic data that may reveal additional insights about the relationship between linguistic complexity of test items and performance gap between the two groups of examinees. However, attempt to use the total score as a proxy of matching ability variable was not good enough because of the complex survey design. Sorting the examinees based on the booklet to derive a total score for matching variable will results in having a smaller sample size per booklet which contradict to Zumbo's (1999) suggestion that sample sizes of 200 are probably appropriate for using logistic regression to detect DIF.

Second, both multiple-choice items and construct-response item were used to investigate DIF on items across examinees to confirm if the differences due to language proficiency may emerge when students are asked to create response or to react to items embedded in more wording. Therefore, we can use the results from this analysis to support the main issues of whether performance difference between the two groups of students can be partly explained by language factors in the assessment, whether the linguistic complexity of test items as a possible source of measurement error or construct-irrelevant variance that can potentially influence the reliability and validity of the test instruments.

It is important to mention that one potential explanation for not many PISA mathematics items was flagged as DIF could be due to the high-quality procedures applied on items selection stage. PISA items and tests undergo several rounds of vigorous review and quality control to ensure the test results have high validity, reliability and most importantly, they are equally fair for both examinees who did or did not speak the language of the test beyond the classroom setting.

However, there are some significant limitations of the study that should be addressed in future research. First, some studies on the translation and adaptation of international tests like PISA and TIMSS, have demonstrated that a large amount of their items suffer from significant problems that limit their interpretive validity. For instance, inaccurate translation and adaptation of the word “wing” in item 80 has contributed to confusion for both groups of students. Second, only 84 mathematics items were included in the analyses because many items were not administered to Malaysian examinees. Also, apart from reporting the hypothesis testing statistics, it would be useful to report the effect size which can be calculated similarly by taken the regression coefficient (R-squared) for the interaction and subtracting the regression coefficient for the total score (Zumbo, 1999). Item can be classified as displaying negligible DIF, moderate DIF or large DIF, according to the criteria established by Jodoin and Gierl (2001). Lastly, it will also be interesting if the result from this study can be used to make inferences to both uniform and non-uniform DIF, to see if the performance differences between the two groups on an item could depend on ability level.

REFERENCES

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement In Education* 14, 219-234.

Carstens, R. & Hastedt, D. (2010). *The effect of not using plausible values when they should be: An illustration using TIMSS 2007 grade 8 mathematics data*. IEA Data Processing and Research Center.

Gennaro, K, (2006). Fairness and test use: The case of the SAT and writing placement for ESL students. *TESOL & Applied Linguistics* 6(2).

Haladya, T. M. & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice* 23 (1), 17-27.

Hauger, J. B. & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing* 8, 237-250.

Kim, Y. H. & Jang, E. E. (2009). Differential functioning of reading sub-skills on the OSSLT for LL and ELL students: a multidimensional model-based DBF/DIF approach. *Language Learning* 59 (4), 825-865.

Klieme, E. & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and DIF in TIMSS. *European Journal of Psychology and Education* 14 (3), 385-402.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.

Mallenbergh, G. J. (1998). Item bias and item response theory. *International Journal of Educational Research* 13, 127-142.

Messick, S. (1989). Validity in R. L. Linn (Ed.), *Educational Measurement, Third Edition* (pp. 13-104). New York: Macmillan.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and DIF for ELL in math tests. *Educational Measurement* 14, 160-179.

Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations? *Journal of Applied Measurement* 10(3).

Pearson, A., & Champagne P.D. (2003), *Subject domain. What is being measured? In NAEP validity studies: An agenda for NAEP validity studies* (pp. 5-12). Washington, DC: National Center for Educational Statistics (NCES Report no. 2003-07).

Pierce, M. E., & Fontaine, L. M. (2009). Designing vocabulary instruction in mathematics. *The Reading Teacher* 63(3), 239-243.

Schmidt, W. H., Cogan, L. S., & McKnight, C. C. (2011). Equality of educational opportunity: Myth or reality in U.S. schooling? *American Educator 34*(4), 12-19.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment* 11(2), 105-106.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias? DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* 58 (2), 159-194.

Sireci, S. G. (1997). Problems and issues in linking assessment across languages. *Educational Measurement: Issues and Practice* 13 (3), 229-248.

Walker, C. M. (2011). What's the DIF? Why DIF analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment* 29 (4), 364-376.

Walker, C. M., Zhang, B., & Sueber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education* 21, 162-181.

Wolf, M., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment 14*(3/4), 139-159.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation,* 31(2-3), 114-128.

Yildirim, H. H. & Berberoglu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing* 9, 108-121.

Zumbo, B. (1999). *A handbook on a theory and methods of DIF: logistic regression modelling 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.

Appendix A

Item Classification for PISA 2012 Mathematics (Malaysia)

▒ Items not administered to Malaysian examinees

Item No	*Unit Item Code*	*Unit Name*	*Item Format 2012*	*Content*	*Context (MS12)*	*Process (MS12)*
1	PM033Q01	P2000 A View Room	Simple Multiple Choice	Space and Shape	Personal	Interpret
2	PM034Q01	P2000 Bricks	Constructed Response Auto-coded	Space and Shape	Occupational	Formulate
3	PM155Q01	P2000 Population Pyramids	Constructed Response Expert	Change and Relationships	Scientific	Interpret
4	PM155Q02	P2000 Population Pyramids	Constructed Response Expert	Change and Relationships	Scientific	Employ
5	PM155Q03	P2000 Population Pyramids	Constructed Response Expert	Change and Relationships	Scientific	Employ
6	PM155Q04	P2000 Population Pyramids	Complex Multiple Choice	Change and Relationships	Scientific	Interpret
7	PM192Q01	P2000 Containers	Complex Multiple Choice	Change and Relationships	Scientific	Formulate
8	PM273Q01	P2000 Pipelines	Complex Multiple Choice	Space and Shape	Occupational	Employ
9	PM305Q01	Map	Simple Multiple Choice	Space and Shape	Societal	Employ
10	PM406Q01	Running Tracks	Constructed Response Expert	Space and Shape	Societal	Employ
11	PM406Q02	Running Tracks	Constructed Response Expert	Space and Shape	Societal	Formulate
12	PM408Q01	Lotteries	Complex Multiple Choice	Uncertainty and data	Societal	Interpret
13	PM411Q01	Diving	Constructed Response Manual	Quantity	Societal	Employ
14	PM411Q02	Diving	Simple Multiple Choice	Uncertainty and data	Societal	Interpret
15	PM420Q01	Transport	Complex Multiple Choice	Uncertainty and data	Personal	Interpret
16	PM423Q01	Tossing Coins	Simple Multiple Choice	Uncertainty and data	Personal	Interpret
17	PM442Q02	Braille	Constructed Response Manual	Quantity	Societal	Interpret
18	PM446Q01	Thermometer Cricket	Constructed Response Manual	Change and Relationships	Scientific	Formulate
19	PM446Q02	Thermometer Cricket	Constructed Response Expert	Change and Relationships	Scientific	Formulate
20	PM447Q01	Tile Arrangement	Simple Multiple Choice	Space and Shape	Societal	Employ

21	PM462Q01	Third Side	Constructed Response Expert	Space and Shape	Scientific	Employ
22	PM464Q01	The Fence	Constructed Response Auto-coded	Space and Shape	Societal	Formulate
23	PM474Q01	Running Time	Constructed Response Manual	Quantity	Personal	Employ
24	PM496Q01	Cash Withdrawal	Complex Multiple Choice	Quantity	Societal	Formulate
25	PM496Q02	Cash Withdrawal	Constructed Response Manual	Quantity	Societal	Employ
26	PM559Q01	Telephone Rates	Simple Multiple Choice	Quantity	Societal	Interpret
27	PM564Q01	Chair Lift	Simple Multiple Choice	Quantity	Societal	Formulate
28	PM564Q02	Chair Lift	Simple Multiple Choice	Uncertainty and data	Societal	Formulate
29	PM571Q01	Stop The Car	Simple Multiple Choice	Change and Relationships	Scientific	Interpret
30	PM603Q01	Number Check	Complex Multiple Choice	Quantity	Scientific	Employ
31	PM800Q01	Computer Game	Simple Multiple Choice	Quantity	Personal	Employ
32	PM803Q01	Labels	Constructed Response Auto-coded	Uncertainty and data	Occupational	Formulate
33	PM828Q01	Carbon Dioxide	Constructed Response Expert	Change and Relationships	Scientific	Employ
34	PM828Q02	Carbon Dioxide	Constructed Response Manual	Uncertainty and data	Scientific	Employ
35	PM828Q03	Carbon Dioxide	Constructed Response Manual	Quantity	Scientific	Employ
36	PM00FQ01	Apartment purchase	Constructed Response Expert	Space and shape	Personal	Formulate
37	PM00GQ01	An advertising column	Constructed Response Manual	Space and shape	Personal	Formulate
38	PM00KQ02	Wheelchair basketball	Constructed Response Expert	Space and shape	Personal	Formulate
39	PM903Q01	Drip rate	Constructed Response Expert	Change and relationships	Occupational	Employ
40	PM903Q03	Drip rate	Constructed Response Manual	Change and relationships	Occupational	Employ
41	PM905Q01	Tennis balls	Complex Multiple Choice	Quantity	Occupational	Interpret
42	PM905Q02	Tennis balls	Constructed Response Expert	Quantity	Occupational	Interpret
43	PM906Q01	Crazy ants	Simple Multiple Choice	Quantity	Scientific	Employ
44	PM906Q02	Crazy ants	Constructed Response Expert	Quantity	Scientific	Employ
45	PM909Q01	Speeding fines	Constructed Response Manual	Quantity	Societal	Interpret
46	PM909Q02	Speeding fines	Simple Multiple Choice	Quantity	Societal	Employ
47	PM909Q03	Speeding fines	Constructed Response Expert	Change and relationships	Societal	Interpret

48	PM915Q01	Carbon Dioxide (CO2) tax	Simple Multiple Choice	Uncertainty and data	Societal	Employ
49	PM915Q02	Carbon Dioxide (CO2) tax	Constructed Response Manual	Change and relationships	Societal	Employ
50	PM918Q01	Charts	Simple Multiple Choice	Uncertainty and data	Societal	Interpret
51	PM918Q02	Charts	Simple Multiple Choice	Uncertainty and data	Societal	Interpret
52	PM918Q05	Charts	Simple Multiple Choice	Uncertainty and data	Societal	Employ
53	PM919Q01	Z's fan merchandise	Constructed Response Manual	Quantity	Personal	Employ
54	PM919Q02	Z's fan merchandise	Constructed Response Manual	Quantity	Personal	Formulate
55	PM923Q01	Sailing ships	Simple Multiple Choice	Quantity	Scientific	Employ
56	PM923Q03	Sailing ships	Simple Multiple Choice	Space and shape	Scientific	Employ
57	PM923Q04	Sailing ships	Constructed Response Expert	Change and relationships	Scientific	Formulate
58	PM924Q02	Sauce	Constructed Response Manual	Quantity	Personal	Formulate
	PM934Q01	London eye	Constructed Response Manual	Space and shape	Societal	Employ
	PM934Q02	London eye	Simple Multiple Choice	Space and shape	Societal	Formulate
	PM936Q01	Seats in a theatre	Constructed Response Manual	Change and relationships	Occupational	Employ
	PM936Q02	Seats in a theatre	Constructed Response Expert	Change and relationships	Occupational	Formulate
	PM939Q01	Racing	Simple Multiple Choice	Uncertainty and data	Societal	Interpret
	PM939Q02	Racing	Simple Multiple Choice	Uncertainty and data	Societal	Interpret
	PM942Q01	Climbing Mount Fuji	Simple Multiple Choice	Quantity	Societal	Formulate
	PM942Q02	Climbing Mount Fuji	Constructed Response Expert	Change and relationships	Societal	Formulate
	PM942Q03	Climbing Mount Fuji	Constructed Response Manual	Quantity	Societal	Employ
59	PM943Q01	Arches	Simple Multiple Choice	Change and relationships	Occupational	Formulate
60	PM943Q02	Arches	Constructed Response Expert	Space and shape	Occupational	Formulate
	PM948Q01	Part time work	Simple Multiple Choice	Quantity	Occupational	Interpret
	PM948Q02	Part time work	Constructed Response Manual	Quantity	Occupational	Employ
	PM948Q03	Part time work	Constructed Response Expert	Quantity	Occupational	Employ
61	PM949Q01	Roof truss design	Complex Multiple Choice	Space and shape	Occupational	Employ
62	PM949Q02	Roof truss design	Complex Multiple Choice	Space and shape	Occupational	Employ

63	PM949Q03	Roof truss design	Constructed Response Expert	Space and shape	Occupational	Formulate
64	PM953Q02	Flu test	Constructed Response Expert	Uncertainty and data	Scientific	Interpret
65	PM953Q03	Flu test	Constructed Response Manual	Uncertainty and data	Scientific	Formulate
66	PM953Q04	Flu test	Constructed Response Expert	Uncertainty and data	Scientific	Formulate
67	PM954Q01	Medicine doses	Constructed Response Manual	Change and relationships	Scientific	Employ
68	PM954Q02	Medicine doses	Constructed Response Expert	Change and relationships	Scientific	Employ
69	PM954Q04	Medicine doses	Constructed Response Expert	Change and relationships	Scientific	Employ
70	PM955Q01	Migration	Constructed Response Manual	Uncertainty and data	Societal	Interpret
71	PM955Q02	Migration	Constructed Response Expert	Uncertainty and data	Societal	Interpret
72	PM955Q03	Migration	Constructed Response Expert	Uncertainty and data	Societal	Employ
	PM957Q01	Helen the cyclist (E)	Simple Multiple Choice	Change and relationships	Personal	Employ
	PM957Q02	Helen the cyclist (E)	Simple Multiple Choice	Change and relationships	Personal	Employ
	PM957Q03	Helen the cyclist (E)	Constructed Response Manual	Change and relationships	Personal	Employ
	PM961Q02	Chocolate	Constructed Response Expert	Change and relationships	Occupational	Employ
	PM961Q03	Chocolate	Simple Multiple Choice	Change and relationships	Scientific	Employ
	PM961Q05	Chocolate	Constructed Response Expert	Uncertainty and data	Occupational	Interpret
	PM967Q01	Wooden train set	Constructed Response Manual	Space and shape	Personal	Employ
	PM967Q03	Wooden train set	Complex Multiple Choice	Space and shape	Personal	Formulate
73	PM982Q01	Employment data	Constructed Response Manual	Uncertainty and data	Societal	Employ
74	PM982Q02	Employment data	Constructed Response Manual	Uncertainty and data	Societal	Employ
75	PM982Q03	Employment data	Complex Multiple Choice	Uncertainty and data	Societal	Interpret
76	PM982Q04	Employment data	Simple Multiple Choice	Uncertainty and data	Societal	Formulate
	PM985Q01	Which car?	Simple Multiple Choice	Uncertainty and data	Personal	Interpret
	PM985Q02	Which car?	Simple Multiple Choice	Quantity	Personal	Employ
	PM985Q03	Which car?	Constructed Response Manual	Quantity	Personal	Employ
	PM991Q01	Garage	Simple Multiple Choice	Space and shape	Occupational	Interpret
	PM991Q02	Garage	Constructed Response Expert	Space and shape	Occupational	Employ

77	PM992Q01	Spacers	Constructed Response Manual	Space and shape	Occupational	Formulate
78	PM992Q02	Spacers	Constructed Response Manual	Space and shape	Occupational	Formulate
79	PM992Q03	Spacers	Constructed Response Expert	Change and relationships	Occupational	Formulate
80	PM995Q01	Revolving door	Constructed Response Manual	Space and shape	Scientific	Employ
81	PM995Q02	Revolving door	Constructed Response Expert	Space and shape	Scientific	Formulate
82	PM995Q03	Revolving door	Simple Multiple Choice	Quantity	Scientific	Formulate
83	PM998Q02	Bike rental	Constructed Response Manual	Change and relationships	Personal	Interpret
84	PM998Q04	Bike rental	Complex Multiple Choice	Change and relationships	Personal	Employ

Appendix B

Analysis for 32 DIF items

Step 1: Using Logistic Regression methods to check whether any mathematics items show DIF among 2 groups of examinees.
Step 2: For items that show DIF, a 2nd run will be conducted while controlling the reading ability.
Step 3: Controlling for SES.

Values in Step 1, Step 2 and Step 3:
Likelihood difference from full model – matching model
Average of 5 values from running the analysis using 5 plausible values

▨ Items that show DIF

Item No	***Item Format 2012***	***Content***	***Context (MS12)***	***Process MS12***	***Step 1***	***Step 2***	***Step 3***
4	Constructed Response Expert	Change and Relationships	Scientific	Employ	8.533153527	15.98803214	-0.388474708
7	Complex Multiple Choice	Change and Relationships	Scientific	Formulate	12.14347102	2.352356754	-0.552874083
9	Simple Multiple Choice	Space and Shape	Societal	Employ	16.22114055	12.21710595	5.547494059
10	Constructed Response Expert	Space and Shape	Societal	Employ	10.07333394	0.193922448	-0.207507497
15	Complex Multiple Choice	Uncertainty and data	Personal	Interpret	15.51764127	0.42502536	-0.88244656
16	Simple Multiple Choice	Uncertainty and data	Personal	Interpret	19.35704269	9.511281552	5.558525717
19	Constructed Response Expert	Change and Relationships	Scientific	Formulate	11.58513004	1.678622429	1.322170437
20	Simple Multiple Choice	Space and Shape	Societal	Employ	9.261568907	1.07800779	1.4319297
25	Constructed Response Manual	Quantity	Societal	Employ	15.86672741	3.008549699	0.929857891
30	Complex Multiple Choice	Quantity	Scientific	Employ	10.49433676	3.248244655	-0.099036902
33	Constructed Response Expert	Change and Relationships	Scientific	Employ	20.35103937	5.21486143	-0.137680793
36	Constructed Response Expert	Space and shape	Personal	Formulate	7.298675553	8.821773224	-0.500551733
44	Constructed Response Expert	Quantity	Scientific	Employ	11.70979985	25.85348004	-1.859162796

45	Constructed Response Manual	Quantity	Societal	Interpret	7.090579716	20.97588742	3.99011539
48	Simple Multiple Choice	Uncertainty and data	Societal	Employ	6.811342848	0.219060504	1.614275301
49	Constructed Response Manual	Change and relationships	Societal	Employ	17.42504748	9.237258813	5.079713434
50	Simple Multiple Choice	Uncertainty and data	Societal	Interpret	8.144095953	10.67766247	-0.269851917
53	Constructed Response Manual	Quantity	Personal	Employ	33.79988449	18.55502397	1.347514214
54	Constructed Response Manual	Quantity	Personal	Formulate	10.41596967	1.800464606	2.298305583
55	Simple Multiple Choice	Quantity	Scientific	Employ	6.190783771	4.02244715	-1.743481837
56	Simple Multiple Choice	Space and shape	Scientific	Employ	11.13183321	5.942734943	7.456033262
59	Simple Multiple Choice	Change and relationships	Occupational	Formulate	8.216291599	6.805668725	-2.301679553
66	Constructed Response Expert	Uncertainty and data	Scientific	Formulate	6.430150666	1.348086473	-0.715912383
67	Constructed Response Manual	Change and relationships	Scientific	Employ	25.38164793	3.134682602	-0.061526821
69	Constructed Response Expert	Change and relationships	Scientific	Employ	10.65039905	1.914848208	-1.712967268
70	Constructed Response Manual	Uncertainty and data	Societal	Interpret	8.116150687	0.717038155	-0.646063783
73	Constructed Response Manual	Uncertainty and data	Societal	Employ	25.03853931	16.20310549	4.178720569
74	Constructed Response Manual	Uncertainty and data	Societal	Employ	153.5834555	18.38663614	4.431097819
75	Complex Multiple Choice	Uncertainty and data	Societal	Interpret	31.16118125	0.724268676	2.139392541
78	Constructed Response Manual	Space and shape	Occupational	Formulate	8.039173839	0.415850537	-0.162604786
80	Constructed Response Manual	Space and shape	Scientific	Employ	22.78577767	2.258412263	-1.739376692
83	Constructed Response Manual	Change and relationships	Personal	Interpret	7.78186733	3.367273365	-1.22778663
	Total				**32**	**15**	**7**