INVESTIGATING ITEM BIAS ON THE PISA 2009 READING ASSESSMENT:

A CASE OF MACAU WITH CHINESE AND ENGLISH VERSIONS

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF
THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF EDUCATION

IN

EDUCATIONAL PSYCHOLOGY

MAY 2017

By

Sok-Han Lau

Thesis Committee:

Seongah Im, Chairperson

Ronald Heck

Lois Yamauchi

Abstract

In recent years, there has been significant increase of regions and countries participating in international large-scale assessments, and this increase is largely due to the extensive information and analysis of the results that are available to the schools, parents, researchers, and educators. Nonetheless, it is questionable whether the results of these international large-scale assessments are reliable, valid, and comparable for different countries. Since most school authorities and educators are interested in using the results of these assessments to enhance the existing school curriculum, educational policy, and program development, the fairness of the assessments and equivalency need to be examined carefully. In this study, a selected sample of the Macau dataset obtained from the Programme for International Student Assessment (PISA) 2009 reading literacy assessment were analyzed to detect potentially biased items using the Mantel-Haenzsel (MH) and item response theory (IRT) methods. Findings indicated that both of the methods commonly identified 5 items that were not working equally across the groups. Limitations and implications were discussed.

Keywords: item bias, differential item functioning, the Mantel-Haenzsel procedure, item response theory, reading literacy, reading assessment, test translation

Acknowledgements

First, I would first like to thank all the members of my committee who have supported me throughout the time I was working on my thesis. I would particularly like to thank Dr. Im, who encouraged me along the way. She consistently reminded me to do the best I could and not to be afraid of making mistakes. The most important of all is that she allowed this paper to be my own work, but guided me in the right direction whenever she thought I needed it. I will always be thankful for her mentorship.

I would also like to thank my family and friends for supporting me throughout my studies in the Master’s program. Thanks also to my fellow classmates who shared their research ideas with me and gave me a helping hand whenever I needed them. Finally, I have to express a special thank you to my parents, who taught me the value of education. I would not be where I am without the love and support of them.

Table of Contents

List of Tables

List of Figures

Investigating Item Bias on the PISA 2009 Reading Assessment: A Case of Macau with Chinese and English Versions

In recent years, there has been an increasing number of countries participating in large-scale international assessments (e.g., PISA, TIMSS, PIRLS). In the Programme of International Student Assessment (PISA) 2000, the number of participating countries was over 40, and it rose to more than 70 in 2015. Around 40 countries participated in the Trends in International Mathematics and Science Study (TIMSS) 1995, and the number increased to over 55 for the TIMSS 2015 assessment. In 2001, the Progress in International Reading Literacy Study (PIRLS) had 35 participated countries and this number increased to over 60 in 2016. These increases may imply that government authorities, researchers and educators are more interested to learn about students' achievement and may have the intention of using the results of the assessments as benchmarks or references to improve school curriculums, programs development and evaluation, and educational policies. Seeing the opportunity to learn more about students' learning and achievement through participating in international large-scale assessments, the Macau government decided to participate in the PISA and PIRLS.

Are the results of these large-scale assessments applicable to the unique culture, language varieties, and education system? As a former Portuguese colony, Macau has maintained two official languages: Chinese and Portuguese. Though Chinese and Portuguese are the official languages, only a small population of people in Macau speaks and uses Portuguese at schools, workplaces, or in their daily lives. On the contrary, English has become more important in Macau as tourism industry has been flourishing for the 10 years. In Macau's non-tertiary education systems (i.e., kindergarten, primary, and secondary), languages of instructions vary according to schools' curricula and needs. In the tertiary education system, some universities require higher proficiency in English and others require higher proficiency in Chinese. The proficiency requirement of a specific language depends on the objectives of the university curriculum. Three languages (i.e., Chinese, English, and Portuguese) and four tongues have coexisted in the Macau school system (Lau, 2007; Shan, 2009; Shan & Ieong, 2009) throughout the history of Macau. In addition to specific language required and used in the non-tertiary and tertiary education sectors, preferences for different forms of English exist as well. Some schools favor British English over American English as the language of instruction and vice versa.

Language varieties are not the only challenge to students and schools in Macau. In the non-tertiary education sector, no standardized tests or other forms of assessment instruments are designed to evaluate students' learning progress or achievement. Each school has its own criteria or standard to determine whether a student has successfully completed all the requirements to graduate from high school. School curricular are not standardized as well. These challenges may be the reasons why the Macau government decided to require secondary schools to participate in international large-scale assessments. Macau has been participating in PISA assessments since 2003 and participated in PIRLS in 2016.

The Macau government has been paying a lot of attention to the Macau reading scores of these large-scale assessments. Reading ability is significant in today's fast growing and changing world. People are constantly reading and searching information on the worldwide web. Educators and most of the general public associate high levels of reading skills to academic success and long-term benefits of employment opportunities and advancement (Hernandez, 2011; Knighton & Bussiere, 2006; Chiswick, 1991). The Organisation for Economic Co-operation and Development (OECD, 2010) defined reading literacy for the PISA 2009 as "understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one’s knowledge and potential, and to participate in society” (p. 23). The definitions were created based on changes in society, economy, and culture.

Reading skills also relate to academic success in other subject areas in the school system, such as mathematics and science achievement. Previous research studies suggest that the development of reading skills begins early in life and these skills are building blocks for more advance reading skills that develop later (Chall, 1983; Snow et al., 1998; Stanovich, 1986). Students who did not develop reading-related skills early in life often encounter reading difficulties when they tackled harder reading materials (Alexander, Entwisle, & Horsey, 1997; Ensminger & Slusarcick, 1992; Juel, 1988). Given Macau’s unique language varieties and educational system, it is unclear to what extent they can depend on the results of these large-scale assessments.

One of the main concerns in large-scale assessments is test bias. Sources of test bias include gender, culture, ethnicity/race, socioeconomic status, language, and special needs (Crane, van Belle, & Larson, 2004; Abbott, 2006; Grossman & Franklin, 1988). In the 1960s, the differential item functioning (DIF) analyses were developed because people were concerned about the unfairness to minority examinees on cognitive ability tests (Angoff, 1993). On a cognitive test, when two groups of examinees possess different levels of knowledge, the group with higher levels of knowledge is more likely to perform better than the other group. Nonetheless, when both groups of examinees possess the same level of knowledge, but have different likelihood of success, results may not be due to differences on ability. DIF occurs when two groups of examinees have the same ability levels, but have different probabilities of obtaining the item correctly (Furr & Bacharach, 2014). In the DIF analysis, a focal group and a reference group are assigned. The focal group consists of examinees for whom researchers are interested in finding out if an item functions differently for them. On the contrary, the reference group consists of examinees to whom focal group is compared. Both uniform and non-uniform DIF can be detected in the DIF analysis (Mellenbergh, 1982). Uniform DIF occurs when the probabilities for the reference group to get a test item correct is consistently higher for the focal group controlling for the ability levels. Non-uniform DIF is present when there is an interaction between ability level and group membership (reference vs. focal group).

The purpose of this study is to investigate possible item bias exist in the PISA 2009 reading literacy assessment of Macau. Common items across the Chinese and English versions of the test were selected prior to the analysis of the testing scores. The first two parts of the study were the identification of possible DIF items using the Mantel-Haenszel (MH) and item response theory (IRT) statistics. The DIF items that were identified from the statistical analysis were further examined. I reviewed some of the DIF items and suggested the possible causes of these DIF items based on my knowledge of the school curricula and culture in Macau.

Literature Review

PISA Assessment

The OECD is an European organization that was established in 1961. The OECD works with governments worldwide to develop policies that will enhance the quality of people's lives. At present, the OECD has 35 member countries and the number of members is expanding. In addition to the member countries, the OECD also works with emerging countries (e.g. Mexico and Turkey), those with emerging economies (the People's Republic of China and India), and other with developing economies in Africa, Asia, Latin America and the Caribbean. In the 1990s, the OECD started working on the PISA assessment because some of the member countries needed reliable and valid assessments to evaluate their students' knowledge and skills and the performance of their education systems. Subsequently, the PISA was introduced in 1997 and the first survey was conducted in 2000. Number of participated countries and economies was 43 in 2000 and 65 in 2012.

PISA is a triennial international survey and the only international education survey that measures the knowledge and skills of 15-year-old students. The age of students are defined by the OECD. At the time of the assessment, students who participate in the survey need to be between 15 years 3 months and 16 years 2 months and have completed a minimum of 6 years of formal education. Areas of assessment include reading, mathematics, and science. Students and school principals complete questionnaires. The responses of both student and school principals questionnaires provide information about students' backgrounds and their learning experiences, schools' backgrounds, the general school systems, and learning environment in each participating region or country. Some countries also administer some optional PISA questionnaires: the computer familiarity questionnaire, the educational career questionnaire and the parent background questionnaire.

The international contractors and all the PISA participants were involved in creating, writing, and reviewing the test questions for potential cultural bias. Pilot studies using the created questions are conducted before these questions are used in the test. Test questions were not designed based on any type of school curricula because the objective of the survey is to assess how well students are able to apply their knowledge into real life situations by the time they complete a certain level of education. Each PISA survey comprises approximately seven hours of test material. The length of the test is two hours and each student will have a different combination of test questions. The test formats are multiple-choice and open-ended questions. PISA tests are graded by assigned test graders who follow the guidelines that are established by the international contractors and the PISA Subject Experts. Participating countries also contribute their suggestions and ideas in the process of grading. Moreover, the test responses are also cross-checked by other experts.

In the PISA assessments, English and French are source versions provided to the participating countries (OECD, 2012, p. 82). These countries were given translation and adaptation guidelines. Participating countries were required to translate all the testing materials and questionnaires into their national languages. National versions of the test were verified against one of the source versions (English or French) by the PISA International Center.

Possible Sources of Test Bias in Large-Scale Assessments

The objectives of international assessments, such as the PISA and PIRLS, are to inform government authorities or school administrators the challenges and difficulties they may be confronting in their school systems (OECD, 2016; IEA, 2016). However, given the diversity of education systems, curricula, languages, cultures, and socioeconomic statuses of the participating countries, the comparability of scores and the use of the scores to develop educational policy are questionable.

Linguistic and cultural characteristics of the assessed countries may influence reading literacy performance (Grisay & Monseur, 2007), and language is part of culture (Jiang, 2000). Thus, it is crucial to examine the equivalence of the reading test that is administered in different languages. Asil and Brown (2016) indicated several factors that may influence reading performance. These include the unique characteristics of different languages, the writing systems, differences in teaching and learning styles across cultures, socioeconomic status, and educational investment. Asil and Brown compared PISA 2009 reading test performances across participating countries. In the measurement invariance analyses, Australia was the reference group and its reading scores were compared with other countries. Findings from the study suggested that the reading items measured the same constructs across countries regardless of the language version of the test. Additionally, almost no uniform DIF appeared when conducted the scalar invariance except New Zealand, Canada, and the United States.

To examine the equivalence of item difficulty of different language versions of an international test, Grisay and Monseur (2007) conducted a study using data from the PISA reading literacy assessment in 2000 and 2001. There were 47 test adaptations for different languages of instruction that were used by the participating countries. Findings showed that countries that used identical versions of the materials with limited adaptations had a lower percentage of DIF items than countries that used translated versions. In addition, the percentage of non-equivalence of item difficulties was higher between Indo-European (e.g. English, French, and Spanish) and non Indo-European languages (e.g. Chinese, Japanese, and Korean) than within Indo-European languages. This finding resonated to Asil and Brown’s suggestions regarding the impact of language characteristics on measurement invariance. Even within a region or country, using different language versions of a test will also lead to the incomparability of the scores (Oliveri & Ercikan, 2011) because of cultural and linguistic differences within regions.

Based on the IRT approach and judgmental reviews, Ercikan (2002) suggested that sources of DIF were related to test adaptations in which differences in meaning, language structure, and format appeared between translated versions of the test. Do these large-scale assessments treat participants fairly? Kankaras & Moors (2014) conducted a study to examine the level of measurement equivalence using the PISA 2009 data set. Findings indicated that the occurrence of uniform DIF for all the 3 assessments (mathematics, science, and reading) across countries and the scores of the Southeast Asian countries were influenced the most by the uniform DIF. Sandilands, Oliveri, Zumbo, & Ercikan (2013) investigated the sources of DIF between English and Spanish versions in PIRLS 2001, and results of the study suggested that the sources of DIF were related to translation, adaptation effects, and the cognitive loadings of items. The cognitive loadings of items was a source of DIF that in favor of both England and the United States when compared to Colombia. On the other hand, adaptation effects did not show any DIF that favored one group or the other.

Test translations can contribute to item bias. One incorrect test item translation may affect the differential functioning of an item across languages (Allalouf, Hambleton, & Sireci, 1999; Ercikan, 1998; Ercikan, 2002; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Gierl & Khaliq, 2001; Solano-Flores, Backhoff, & Contreras-Nino, 2009). Thus, equivalence or comparability is a major threat to the validity of the test in any international assessment. When participants in different countries take the same test, but in different languages, comparison of the results are questionable because some constructs of the test may not be able to be inferred from the measurements (Ercikan, 1998). Based on test theory, equivalence is defined as using two or more versions of a test and these versions are interchangeable with each other (Arffman, 2010). It also indicates a relationship between an original text and a translated text in which the translated text is considered as a translation of the original text (Kenny, 1998, & Pym, 1995; as cited in Arffman, 2010). Translation of the text is one possible cause of non-equivalency. According to Ercikan, when translations are not done properly, they can influence the psychometric properties of tests. Ercikan (1998) conducted a study using DIF procedures to investigate test items equivalence and scores comparability from tests in different languages. Data were collected from the 1984 International Association for the Evaluation of Educational Achievement (IEA) science tests in English and French, with 5,543 English-speaking students and 2,348 French-speaking students. In the study, 70 common items in both language versions were analyzed. Findings in this study indicated that 8 out of the 18 DIF items were related to translation problems. The government or authorities should interpret the testing scores critically. Test translation and adaptation also appear to be problematic in large-scale assessments.

Applying the theory of test translation error, Solano-Flores, Contreras-Nino, and Backhoff (2013) investigated the measurement of translation error in PISA 2006 science and mathematics items from the Mexican, Spanish language version. In this study, a translation review panel was convened for in the test translation review. Error coding procedures and the examination of the Pearson correlations were performed. Findings suggested that most of the translation errors did not cause test bias, but they might have threatened the validity of the test items. Moreover, correlations between number of different translation error dimensions and item difficulty for mathematics items were higher than the correlations between number of different translation error dimensions and item difficulty for science items. Stubbe (2011) conducted a study using data from PIRLS 2006 and 2007. In this study, Austria, Germany, Luxembourg, and the German-speaking community of Belgium used three different German versions of the PIRLS 2006 reading assessment. Findings indicated that around 74% of the overall items showed a significant DIF. However, when comparisons were performed in pairs (countries that used identical translations), there were less DIF items. Another large-scale study conducted by Budgell, Raju, & Quartetti (1995) using item response theory (IRT) procedures also suggested that DIF items occurred when two different versions (English and French) of the test were analyzed for equivalence.

Using different school and home languages other than the test language in a region or country can hinder students’ test performance on large-scale international assessments. In the case of Macau, students generally learn both Chinese and English at school. However, the written and spoken languages of Chinese may not be the same for these students when they are outside the classroom. They may be learning Mandarin at school, but speak Cantonese at home. Thus, these students may have some disadvantages over other students for whom home and test languages are the same. A study conducted by Ercikan, Roth, Simon, Sandilands, and Lyons-Thomas (2014) examined the effect of home language (French or not French), school language (French) and the test language (French). In this study, three Canadian French versions of the assessment for two linguistic groups in Canada were administered. Findings indicated that there were more DIF items for minority Francophone students, and these students were identified as students who received French instruction at school, but did not speak French at home. The mismatch of test language and home language not only influenced the fairness of the assessment, but also influenced the overall reading achievement in large-scale assessments. To examine test language effect, Soh (2014) conducted a study using PISA 2009 reading assessment data. Results of this study showed that the mean for the national reading sample underestimated the performance of test takers who spoke the test language at home and overestimated test takers who did not speak the test language at home.

Due to differences in teaching approaches and curricula, item format (e.g. multiple choice and open-ended response) in large-scale assessments can also be possible sources of item bias. Multiple-choice items might be more familiar to students in some countries, but not others, while students from developing countries might have disadvantages responding to constructed response items (Grisay & Monseur, 2007). Daneman and Hannon (2001) found that students could answer multiple-choice items without comprehending the reading passage (Rauch & Hartig, 2010). Shohamy (1984) conducted a study to examine the effect of testing methods for reading comprehension on reading performance. Findings in this study suggested that different item formats produced different levels of difficulty for test takers, with multiple-choice items being easier to respond to, compared to open-ended response items. Grisay and Monseur (2007) found that low-performing countries performed better with multiple-choice items, while most of the Asian countries performed better in constructed response items.

Another possible cause of item bias is item dependency (Grisay & Monseur, 2007). Item dependency may be related to cultural differences and level of difficulty of the reading texts. In Arffman's (2010) study, one of the equivalence problems was caused by cultural differences. For example, one of the narrative texts was about Mississippi in America. Because of the content and the context, it was more difficult and perhaps less interesting for the Finnish to comprehend the reading than for Americans. Previous research also supported the notion that items with familiar content and special interest might favor one group of examinees over others (Schmitt, 1988).

Methods to Detect Item Bias

There are various approaches to detect potential item bias using statistics. Methods using item response theory (IRT) employ the Wald Test (Cai, 2012; Cai, Thissen, & du Toit, 2011; Langer, 2008), the Likelihood Ratio Test (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993), and other types of tests. Popular non-IRT methods include the Mantel-Haenszel (MH) Test (Holland & Thayer, 1988), and the method using Logistic Regression (Swaminathan & Roger, 1990).

Item Response Theory (IRT) are mathematical models in which person and item parameters are estimated (Embretson & Reise, 2000). In IRT, each item of a test is assumed to measure the underlying latent trait, or ability (θ). The probability of examinees' correct responses to the *i*th item ($P_i(\theta)$) increases as the level of the ability (θ) increases.

The equations for the logistic IRT Models:

1 parameter logistic (1PL) model:

$$P(Y = 1|\theta) = \frac{e^{1.7(\theta - b_i)}}{1 + e^{1.7(\theta - b_i)}}$$

2 parameter logistic (2PL) model:

$$P(Y = 1|\theta) = \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}$$

3 parameter logistic (3PL) model:

$$P(Y = 1|\theta) = c_i + (1 - c_i) \cdot \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}$$

In these models, theta(θ) is the latent trait score of examinees, *a* is the discrimination parameter, *b* is the difficulty parameter, and *c* is a guessing parameter.

IRT method using the improved Wald Test is the latest psychometric method to detect DIF. Lord’s Wald test (1980) is a statistic that compares parameter estimates for an item between the reference and focal groups that divided by the standard error of their difference (Woods, Cai, & Wang, 2012). Lord’s (1980) original Wald Test for DIF examined item parameters in each group separately and the metric θ is subsequently equated. In 2008, Langer introduced a two-stage equating procedure (Wald-2) that improved Lord’s original Wald Test by estimating the covariance matrix using the supplemented expectation maximization (SEM) algorithm. Cai, Thissen, and du Toit (2011) later introduced a one-stage equating procedure (Wald-1) using the SEM algorithm. Another IRT method to detect the DIF is the Likelihood Ratio Test (LR). The IRTLR tests “compare two-group item response models with varying constraints to evaluate whether the response function for a particular item differs for the reference and focal groups” (Woods, 2008).

The equations for the Wald test:

For one parameter,

$$x^2 = \frac{\left(b_f - b_r\right)^2}{\sigma^2_{(b_f - b_r)}}$$

For two or multiple parameters,

$$\mathcal{X}^2 = \mathrm{V}' \, \Sigma^{-1} \, \mathrm{V}$$

$\mathrm{V} = \begin{pmatrix} a_f & -a_r \\ b_f & -b_r \end{pmatrix}$, $\sum$ is the 2 x 2 diagonal covariance matrix of differences of a and b estimates

The Mantel-Haenszel (MH) Test is one of the popular non-IRT methods to detect the DIF. Mantel and Haenszel (1959) method is used for determining whether two variables are independent of one another while conditioning on a third variable. The overall ability or the total score controls for the third variable. For each ability level, a 2 x 2 chi-square contingency table is arranged. In the tables, number of correct and incorrect items for each group at each ability level is counted. The calculation of the MH test uses the odds ratio (α_{MH}) which denotes a linear relationship between the row and the column variables in the tables. The ratios are the proportions of 1 and 0 responses in reference and focal groups given the score of k. Penfield (2003) suggested Breslow-Day (BD) statistic that tests homogeneity of odds ratios for stratum across the score distribution as a way to detect non-uniform DIF. As non-uniform DIF increases, odds-ratios become more heterogeneous. Another non-IRT method is Logistic Regression (Swaminathan & Roger, 1990). Logistic regression is a technique for making predictions about a binary variable using categorical and continuous predictors. This technique can detect both uniform and non-uniform DIF. In logistic regression, model fit is tested first. The loglikelihood values of different models are compared. Then, a chi-square test is used to examine the presence of uniform and non-uniform DIF on the item of interest by testing each term included in the model. Uniform DIF occurs if the group difference in performance on the item is not equal to 0. The presence of non-uniform DIF exists when the interaction between group and ability is not equal to 0.

The equation for the Mantel-Haenszel test:

$$MH = \frac{(|\sum_k a_k - \sum_k E(a_k)| - 0.5)^2}{\sum_k var(a_k)} \sim \chi^2(1)$$

$$\alpha_{MH} = \sum_k \alpha_k = \frac{\sum_k a_k d_k / n_k}{\sum_k b_k \, c_k / n_k}$$

$\alpha_k = \frac{p_{rk}/q_{rk}}{p_{fk}/q_{fk}} = \frac{a_k d_k}{b_k c_k}$ is the odds ratio and p_{rk}, p_{fk}, and q_{rk},q_{fk} are proportions of 1 and 0 responses in reference and focal groups given the score of k. *a* is the number of examinees who scores the item correctly in the reference group (e.g. male). *b* is the number of examinees who scores the item incorrectly in the reference group. *c* is the number of examinees who scores the item correctly in the focal (e.g. female), and *d* is the number of examinees who scores the item incorrectly in the focal group.

The equation for the Breslow-Day tests:

$$\mathrm{BD} = \frac{[\sum_{k=1}^{K} X_k(a_k - \ A_k]^2}{\sum_{k=1}^{K} X_k^2 \, V(a_k) - \frac{\left[\sum_{k=1}^{K} X_k V(a_{k)}\right]^2}{\sum_{k=1}^{K} V(a_k)}}$$

$$V(a_k) = \left[\frac{1}{A_k} + \frac{1}{B_k} + \frac{1}{C_k} + \frac{1}{D_k}\right]^{-1}$$

In this equation, X_k is the value of the *k*th level of the stratifying variable. A_k is the expected frequency of a_k.

Other than using statistical methods, judgmental reviews can also be used to identify DIF. Back-translation is a famous judgmental method to evaluate the equivalence of two language forms (van de Vijver & Leung, 1997). In this method, the original language test is first translated into the target language, and then back-translated into the original language by a different translator. After that, a reviewer or committee of reviewers compares the original version and back-translated version to assess the equivalence of the languages in the original and target versions. Because of the limitations in the back-translation method, Brislin (1986) proposed a modified back-translation method. In this modified method, the original language test is translated into the target language by the test developer. Then, the target language test is back-translated into the original language by two translators separately and independently. The two back-translations and the original language test are then compared for equivalence. Translation problems will arise if there are discrepancies between the two back-translations and the original language. Judgmental review procedures can also be conducted after identifying the DIF items using statistical methods as a way to cross-validate the outcomes and enhance the translation process in the future (Ercikan, 2002; Budgell et al.).

In addition to the use of judgmental reviews to detect item bias, the comparative linguistic text analysis (Brinker, 1992; as cited in Arffman, 2010) is another method that can be used. In this method, the source text and the translation text are compared for different linguistic features (e.g. syntax, lexis) and levels (e.g. semantics, pragmatics, structures). The aim of the comparison is to locate and examine possible linguistic elements on the translation text that seemed to be non-equivalent in difficulty to the source text. Then, the non-equivalent items are calculated quantitatively in different steps. The first step is to calculate the non-equivalences in the data. The next step is to examine the proportion of the number of the different types of non-equivalences to the total number of non-equivalences in each text. The final step was to compare the equivalence of difficulty between the source and translated texts based on theories and the frequencies calculation in the previous steps.

Research Questions

Based on the literature review, this study examined the following questions:

1) Appling the MH method, which items function differently due to different languages used for each version of the PISA 2009 reading assessment?
2) Applying the IRT method, which items function differently due to different languages used for each version of the PISA 2009 reading assessment?
3) Are DIF items found by the two different methods consistent?
4) If items are found to function differently due to different languages used for each version of the reading assessment in PISA 2009, can these differences be explained by different linguistic characteristics or translation methods?

Method

The PISA Reading Assessment and Items

In this study, the reading test of PISA 2009 was selected because the reading framework and the instruments were reexamined and reviewed as the major domain as to better understand the process of reading and fit the changes in the world (OECD, 2010a, p.20). With the new revision of the reading test, more information is available to investigate possible sources of item biases. Based on different estimation methods, the overall reliability coefficients of the PISA 2009 reading test ranged from 0.84 to 0.921 (OECD, 2012). The reliability coefficient of the Macao-China sample was 0.89 (OECD, 2012).

In the reading test, there were a total of 37 reading passages with 131 questions. Students took different sets of combinations of questions on the test day. Each reading passage had one single question format or a combination of question formats: multiple choice, complex multiple choice, open constructed response, closed constructed response, and short responses. Scoring for multiple choice and closed constructed response were scored dichotomously (i.e., correct or incorrect) while open constructed response, short response, and complex multiple choice were either dichotomously scored or polytomous scored using partial credits.

A total of thirteen booklets were used in the reading test in Macau, and each booklet was composed of three language versions (Chinese, English, and Portuguese). Because the Portuguese version of the test comprised of only a small number of students and this language version was not a focus of this study, the Portuguese version of the test in each booklet was omitted. In this study, only reading items that were administered across Chinese and English versions were examined and analyzed.

To search for common items across Chinese and English versions of the test, I first identified all these items in each booklet. Then, I used the information to sort the maximum number of common items across the two language versions and the 13 booklets. Seven possible reading passages in different combination of booklets that were common across the Chinese and English versions were selected. Booklets 1, 3, 4, and 6 with reading passages on Acne Vulgaris (4 items), Mobile Phone (4 items), Telecommuting (3 items), and The Play and the Thing (4 items) were selected for the analysis of this study. Fifteen common items were identified.

Participants

In the initial screening, all the subjects who completed booklets 1 to 13 were considered. The Macau reading scores data set comprised of 5,952 students from 45 schools, with 2,941 females and 3011 males. Among them, 5,316 students took the Chinese version and 601 students took the English version. To make results comparable, only students who took the same test items across the English and Chinese versions were selected for analysis. After the 15 common items were identified across booklets and the two language versions, there were 1,642 students in the Chinese version and 180 students in the English version.

Prior to the statistical analysis, missing values were examined and participants who had missing values in two or more reading passages at the same time were deleted from the sample. After the deletion, there were 1,617 students in the Chinese version and 171 in the English version. This sample was used for further analysis.

Random sampling procedures were also conducted after the omission of missing values. To ensure the comparability of the sample to be used for further analysis, three iterations using 50% and 40% for selection were conducted and the results were compared. After conducting the random sampling procedure, the sample for the analysis of this study was selected. In this sample, there were 772 students in the Chinese version and 171 students in the English version.

Data Analysis

In this study, the DIF analysis using the two groups were performed with the non-parametric procedures using the Mantel-Haenszel (MH) and the Breslow-Day (BD) tests and Wald tests in item response theory. The MH method is used to detect uniform DIF. In this method, each item is studied separately. Examinees are divided into a focal (i.e. English) and a reference (i.e. Chinese) group based on the language version of the test they take. Next, for each ability level, a chi-square contingency table presents the number of examinees who scored *n* points in each cell. In the MH DIF procedure, I examined whether the MH statistic was significant and then followed by the examination of the size of the common odd ratio.

Additionally, the Breslow-Day chi-square statistic was examined. The BD statistic has shown to be effective to detect nonuniform DIF. DIFAS 5.0 (Penfield, 2012) was used for this analysis.

Followed by the MH method, the item response theory method using the improved Wald Test were performed. First, the Wald-2 test was conducted. In this test, anchor items were not identified. In the first stage of the test, the reference group (i.e. Chinese) mean and standard deviation were fixed to 0 and 1 as to identify the scale; the mean and standard division of the focal group (i.e. English) were estimated; all the item parameters were constrained equally between groups. In the second stage of the test, the mean and standard deviation of the focal group were fixed to the values obtained from the first stage; all item parameters were free to vary between groups. Wald-1 Test was conducted after the Wald-2 test.

In the Wald-1 test, anchor items were specified based on the testing in the Wald-2 Test. The mean and standard deviation of the reference group were fixed to 0 and 1; the mean and standard deviation of the focal group were estimated concurrently with the estimation of the item parameters. IRTPRO for Student (IRTPRO, 2015) and SPSS 24.0 (SPSS, 2016) were used for this analysis.

Results

To answer the first question of the research question, the non-parametric procedure was used for the analysis. DIF statistics using the Mantel-Haenszel (MH) and Breslow-Day (BD) statistics are reported in Table 1. The MH statistic follows a chi-square distribution with one degree of freedom and critical values are 3.84 with α at .05 and 6.63 with α at 0.01. Findings using the MH CHI statistic indicate that Acne1 (item 5 = 36.02) had values a lot greater than the chi-square critical value. Tele2 (item 14 = 14.07), Mobile4 (item 4 = 13.52), and Play2 (item 10

= 10.64) also had values greater than the chi-square critical value. These four items were identified as DIF items and were statistically significant.

Another statistical method to identify possible DIF items is the Standardized Mantel-Haenszel Log-Odds Ratio (LOR Z). To calculate the LOR Z, I divided the Mantel-Haenszel log-odds ratio (MH LOR) by the estimated standard error (LOR SE). Values obtain from the LOR Z that are less than -2 or greater than 2 suggest the presence of potential DIF items. Results using this statistic showed that Play2 (item 10 = 3.27), and Tele2 (item 14 = 3.74) had values greater than 2 while Acne1 (item 5 = -6) and Mobile4 (item 4 = -3.70) had values less than -2. While the MH statistic is a robust procedure to detect the presence of uniform DIF, the Breslow-Day Test of Trend in Odds Ratio (BD) has been shown to be effective at detecting non-uniform DIF (Prieto-Maranon, Aguerri, Galibert, & Attorresi, 2011). The null hypothesis with BD tests state that there is no odds ratio homogeneity. When odds-ratios become more heterogeneous, chi-square value increases and in which the non-uniform DIF increases. The BD follows a chi-square distribution with one degree of freedom (Breslow & Day, 1980; Penfield, 2003). Critical values are 3.84 with p = 0.05 and 6.63 with p = 0.01. Results from the BD statistics showed that Play2 (item 10 = 3.26) and Tele2 (item 14 = 3.09) were weak non-uniform DIF items. The interactions between the Chinese and English groups were small.

Table 1. The Mantel-Haenszel and Breslow-Day Statistics

Items	MH CHI	LOR Z	BD
Mobile1	0.26	-0.60	0.71
Mobile2	0.64	-0.89	2.33
Mobile3	3.89*	2.09	0.83
Mobile4	13.52**	-3.70	1.99
Acne1	36.03**	-6.00	0.46
Acne2	1.65	1.37	0.16
Acne3	2.42	-1.65	0.17
Acne4	3.03	1.85	0.02
Play1	1.74	-1.53	0.89

Play2	10.65**	3.27	3.26
Play3	4.45*	2.17	1.61
Play4	1.39	1.29	2.11
Tele1	0.15	-0.49	0.02
Tele2	14.07**	3.74	3.09
Tele3	2.99	-1.84	0.61

Note: *p<.05; **p<.01

Using the IRT approach, the improved Wald Test was conducted to answer the second research question of this study. For the first iteration of the Wald Test, the Wald-2 Test was conducted. Anchor items were not designated in the Wald-2 test. The chi-square values in the last column in Table 2 indicate the degree of interaction between the focal and reference groups. Results in Table 2 (second column) did not show the presence of non-uniform DIF. However, six items (Acne1, $\chi^2_{\text{uniform}} = 27.70$, Mobile4, $\chi^2_{\text{uniform}} = 11.3$, Tele2, $\chi^2_{\text{uniform}} = 10.6$, Play2, $\chi^2_{\text{uniform}} = 10$, Play3, $\chi^2_{\text{uniform}} = 4.1$, Play1, $\chi^2_{\text{uniform}} = 4$) were DIF candidates. With six potential uniform DIF items identified from the Wald-2 test, another iteration using Wald-1 Test was examined.

Table 2. Wald-2 Test

Items	**Total χ^2**	**$\chi^2_{\text{non-uniform}}$**	**χ^2_{uniform}**
Mobile1	1.50	1.00	0.50
Mobile2	3.30	1.70	1.60
Mobile3	3.60	0.50	3.00
Mobile4	13.00	1.70	11.30**
Acne1	28.20	0.40	27.70**
Acne2	1.50	0.10	1.40
Acne3	3.20	0.40	2.80
Acne4	3.00	0.00	3.00
Play1	4.10	0.10	4.00*
Play2	11.20	1.20	10.00**
Play3	4.50	0.40	4.10*
Play4	3.60	2.90	0.70
Tele1	1.30	0.00	1.30
Tele2	11.40	0.70	10.60**
Tele3	5.60	1.40	4.20

Note: *p<.05; **p<.01

As six candidate items were identified from the prior testing using Wald-2 test, the anchor set consisted of nine items. In Table 3, only five items were identified as non-uniform DIF items at this stage of the iteration and no uniform DIF item were identified. The potential non-uniform items were: Acne1, χ^2 $_{\text{uniform}}$ = 17.30, Mobile4, χ^2 $_{\text{uniform}}$ = 9.9, Tele2, χ^2 $_{\text{uniform}}$ = 9.5, Play2, χ^2 $_{\text{uniform}}$ = 8.5, and Play3, χ^2 $_{\text{uniform}}$ = 4.2. Since the chi-square value of one of the items was very close to the critical value, another iteration using Wald-1 test waa conducted.

Table 3. Wald-1 Test

Items	**Total χ^2**	**χ^2non-uniform**	**χ^2uniform**
Mobile4	11.10	1.20	9.90**
Acne1	17.50	0.20	17.30**
Play1	1.90	0.10	1.80
Play2	9.00	0.50	8.50**
Play3	4.60	0.40	4.20*
Tele2	9.80	0.40	9.50**

Note: *p<.05; **p<.01

For the second iteration using Wald-1 Test, 10 anchor items and five candidate items were used. In Table 4, results indicated that non-uniform DIF items were not present, but five potential uniform DIF items were present in the sample. The five uniform DIF items were: Acne1, χ^2 $_{\text{uniform}}$ = 16.30, Tele2, χ^2 $_{\text{uniform}}$ = 9.9, Mobile4, χ^2 $_{\text{uniform}}$ = 9.6, Play2, χ^2 $_{\text{uniform}}$ = 8.9, and Play3, χ^2 $_{\text{uniform}}$ = 4.7.

Table 4. Results of Wald-1 Test in Second Iteration

Items	Total χ^2	χ^2non-uniform	χ^2uniform
Mobile4	10.70	1.10	9.60**
Acne1	16.40	0.20	16.30**
Play2	10.30	1.40	8.90**
Play3	5.80	1.00	4.70*
Tele2	10.90	1.00	9.90**

Note: *p<.05; **p<.01

In Table 5, the mean and variance were set to 0 and 1 for the reference group and the mean and variance of the focal group were estimated and equated for the scale of θ.

Table 5. Mean and Variance of the Chinese and English group

Group	μ	σ
Chinese	0.00	1.00
English	0.29	0.92

Table 6 shows the item parameters for the 2PL model for the Chinese and English groups after three iterations.

Table 6. Item parameters for the Chinese and English Groups After Three Iterations

	Item Discrimination		Item Difficulty	
Items	Chinese	English	Chinese	English
Mobile1	0.85	0.85	0.72	0.72
Mobile2	1.58	1.58	0.26	0.26
Mobile3	0.87	0.87	-0.28	-0.28
Mobile4	0.65	0.38	0.80	2.81
Acne1	1.19	1.05	-0.48	0.49
Acne2	1.11	1.11	-0.20	-0.20
Acne3	0.92	0.92	-0.92	-0.92
Acne4	1.09	1.09	0.66	0.66
Play1	1.41	1.41	3.15	3.15
Play2	1.00	1.44	0.55	-0.06
Play3	0.67	0.96	1.27	0.49
Play4	0.88	0.88	1.12	1.12
Tele1	1.23	1.23	0.41	0.41
Tele2	0.88	1.21	0.74	0.01
Tele3	1.07	1.07	0.59	0.59

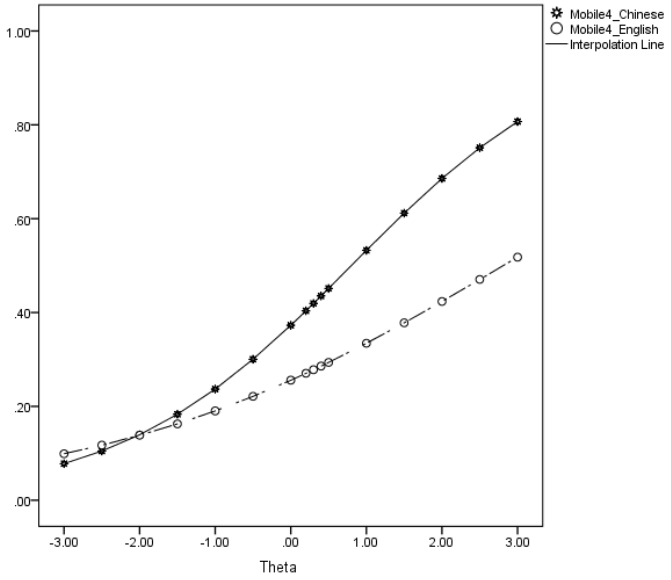

Figure 1. Item Characteristic Curve of Chinese & English Groups (Mobile 4)

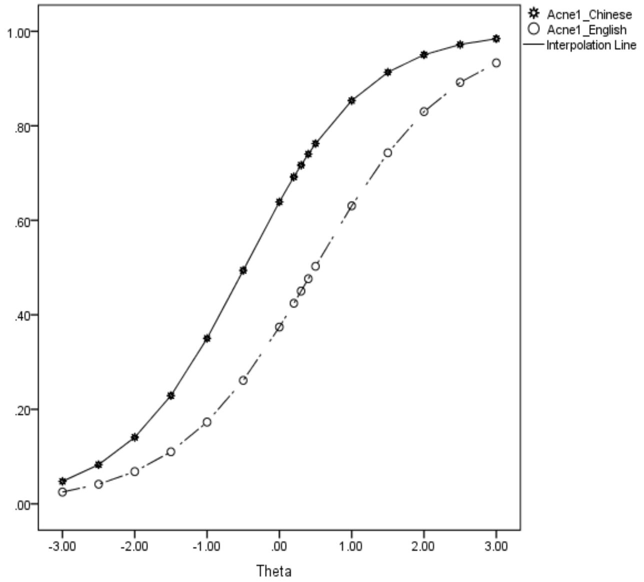

Figure 2. Item Characteristic Curve of Chinese & English Groups (Acne1)

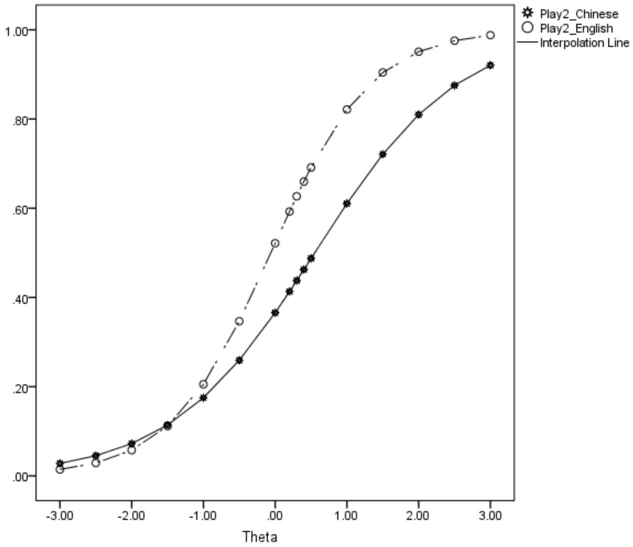

Figure 3. Item Characteristic Curve of Chinese & English Groups (Play2)

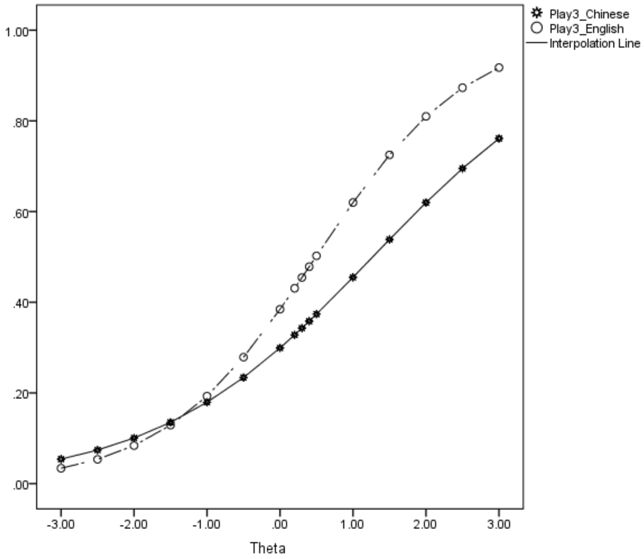

Figure 4. Item Characteristic Curve of Chinese & English Groups (Play 3)

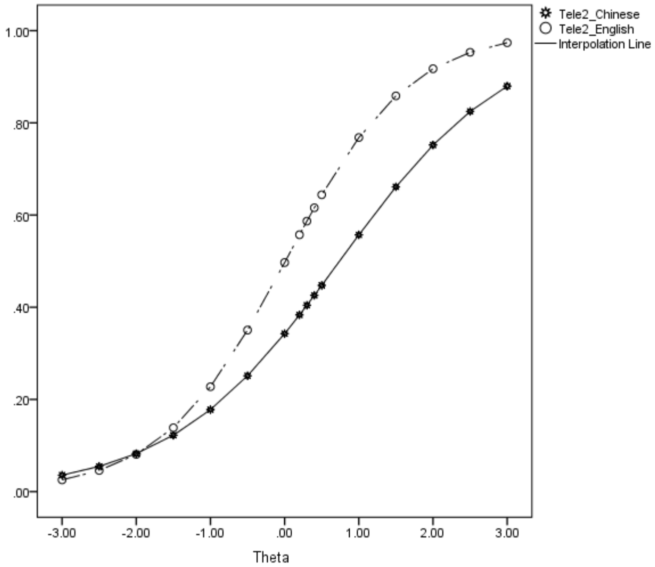

Figure 5. Item Characteristic Curve of Chinese & English Groups (Tele2)

Based on the Item Characteristics Curves (ICCs) (Figures 1 – 5) from above, three out of five DIF items (Play2, Play3, and Tele2) appeared to be more in favor of the English group.

Comparing the two statistical methods, similar results are shown in Table 7. Both methods identified the same DIF items. Acne1 had the most problematic items among all the DIF items, followed by the item Tele2. In responding to the third research question, the two statistical methods showed the same results.

Table 7. DIF items from two different statistical methods

	MH	BD	IRT
Mobile4	high DIF (Chinese)	no non-uniform DIF	in favor of Chinese group
Acne1	high DIF (Chinese)	no non-uniform DIF	in favor of Chinese group
Play2	high DIF (English)	no non-uniform DIF	in favor of English group
Play3	high DIF (English)	no non-uniform DIF	in favor of English group
Tele2	high DIF (English)	no non-uniform DIF	in favor of English group

Though some items were found to function differently for the two groups of examinees, quantitative methods alone are not sufficient to evaluate the fairness of tests. Judgmental reviews are necessary to complement the statistical results. With limited resources and access to the source version and translated version of the tests, further investigation into the linguistic features were not feasible in this study.

Conclusions

In this study, I investigated possible item bias in the PISA 2009 reading assessment of Macau, employing statistical techniques from both non-IRT and IRT approach. Applying the non-IRT approach, the MH-method, five DIF items were identified. All these items showed uniform DIF and non-uniform DIF was not found. Applying the improved Wald Test in the IRT approach, five uniform DIF items were identified. The existence of DIF items did not provide information on the cause of the DIF (Penfield & Camilli, 2007). The causes of DIF may be related to test format, test translation and adaptation, and test language (Allalouf, Hambleton, & Sireci, 1999; Cawthon, Leppo, Carr, & Kopriva, 2013; Kobayashi, 2002; Stevenson, Heiser, & Resing, 2016). One way to examine the possible causes of DIF is to analyze the source version of the reading assessment and its adapted version. Although this step was proposed in this study, further analysis using these reading passages was not feasible because not all the reading passages (source and adapted versions) in PISA 2009 were available for the public.

Discussions

After reviewing the content of the reading passages and test formats, I identified a few issues that may be the causes of the DIF items based on the English version of the test. One of the issues was the content of the reading passages. The themes of the reading passages were on mobile phone safety, acne vulgaris, a play, and telecommuting. Though the design of the PISA assessments is not from a particular type of curricula, students' exposure to different types or genres of reading are related to their school curricula. The curriculum of the English medium schools mostly follow the European education Framework. Students who are in the European Education Framework may be more familiar with certain reading genres, such as the reading passage from a play written by the Hungarian dramatist Molnár (Appendix). The results of the first part of the study showed that two items from The Play's The Thing were in favor of the English group. Additionally, students from the English medium schools may have more opportunities to participate in international assessments, such as the Cambrige IGCSE. Therefore, they are more familiar with different test formats. Students who attend Chinese medium schools may perform better in the reading literacy assessment when the content of the reading passages involves technical vocabulary words. For example, the reading passage on Mobile Phone Safety. Students from Chinese medium schools may understand the content better because they learn the concepts in other classes or from people around them. The curricula of Chinese medium schools tend to focus on mathematics and science subjects as early as students start their secondary school education.

As mentioned earlier in this paper, a wide language varieties exist in Macau. Even within the non-tertiary education system, three languages exist without the inclusion of other dialects. However, we only investigated item bias across two different language versions of the assessment in this study. Possible DIF items may appear within the Chinese version of the assessment since examinees may speak a different dialect at home or at school that is different than the test language. This issue also applies to the examinees who take the English version of the test.

Limitations

Age and grade level, which were not part of the investigation may have also played crucial roles in the results. PISA test assesses students who are 15-years-old regardless of their grade levels. In-grade retention rates of Macao were one of the highest among the OECD countries, with 6 percent and 12 percent repeaters at primary and secondary education respectively (UNESCO, 2011). It is possible that grade retention could affect the average score of a region if students are from a range of grade levels because some of them studied different grade levels more than once. In PISA 2009, the average score of Macau was 487 and the International average score was 493. Another issue related to the grade retention rates in Macau. Students who were retained and did not perform well academically demonstrated low self-esteem (Dawson, 1998; Jimerson, 2001; Thomas, 1992). Are these students motivated to take a low stake test? If they are not motivated, will they put in any effort to do the best they can?

Limited research has investigated the reading domain in PISA assessments (Asil and Brown, 2016). One reason could be the inaccessibility to the reading items. Reading literacy was highly influenced by the cultural, linguistic, and curricular characteristics of participating countries (Grisay & Monseur, 2007). Without being able to examine the reading items, it is not possible to identify specific issues that affected the testing scores. In this study, we can only conclude that DIF items were identified but their causes cannot be identified.

The Macau government has been actively involved in promoting students' academic success in the non-tertiary education. With limited resources to conduct educational research within the Department of Education, the government has been collaborating with the Educational Testing and Assessment Research Centre at the University of Macau. This research center has already conducted numerous research studies with the PISA assessments (Cheung, Mak, Sit, & Soh, 2016; Cheung, 2015). The most recent study conducted by Cheung et al. (2016) was on student reading engagement. Another study conducted by Cheung (2015) was on the factors that affected students’ mathematical performance in Macau. However, to my knowledge, they have not investigated possible item bias across language versions of the assessments. This study can inform the Macau government, educators, and school principals that the results of large-scale assessments can only be used as a reference. They need to interpret and use the results carefully.

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*(1), 7-36.

Alexander, K. L., Entwisle, D. R., & Horsey, C. S. (1997). From first grade forward: Early foundations of high school dropout. *Sociology of Education*, 87-107.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185-198.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 3-23). Hillside, NJ: Lawrence Erlbaum.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educationl Research*, *54*(1), 37-59.

Asil, M., & Brown, G. T. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, *16*(1), 71-93.

Brislin, R. N. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field Methods in Cross Cultural Research* (pp. 137-164). Beverly Hills, CA: Sage Publications.

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

Cai, L. (2012). flexMIRT: *Flexible multilevel item factor analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: *Flexible, multidimensional, multiple categorical IRT modeling* [Computer software] Chicago, IL: Scientific Software International.

Cawthon, S., Leppo, R., Carr, T., & Kopriva, R. (2013). Toward accessible assessments: The promises and limitations of test item adaptations for students with disabilities and English Language Learners. *Educational Assessment*, *18*(2), 73-98.

Chall, J. (1983). *Stages of reading development.* New York: McGraw-Hill.

Cheung, K.C., Mak, S.K., Sit, P.S., & Soh, K.C. (2016). A typology of student reading engagement: Preparing for response to intervention in the school curriculum. *Studies in Educational Evaluation*, *48*, 32-42.

Cheung, K.C. (2015). Factors affecting mathematical literacy performance of 15-year-old students in Macao: The PISA perspective. In B. Sriraman, J. Cai, K.H. Lee , L. Fan, Y. Shimuzu, C.S. Lim, & K. Subramaniam, (Eds.) *The first sourcebook on Asian research in mathematics education: China, Korea, Singapore, Japan, Malaysia and India* (pp. 91 -110). Charlotte, NC: Information Age Publishing.

Chiswick, B. R. (1991). Speaking, reading, and earnings among low-skilled immigrants. *Journal of Labor Economics*, *9*(2), 149-170.

Crane, P. K., Belle, G. V., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine*, *23*(2), 241-256.

Dawson, P. (1998). A primer on student grade retention: What the research says. *NASP Communique*, *26*(8), 28-30.

Embretson, S., & Reise, S. (2000). *Item Response Theory for Psychologists.* Lawrence Erlbaum Associates Publishers.

Ensminger, M. E., & Slusarcick, A. L. (1992). Paths to high school graduation or dropout: A longitudinal study of a first-grade cohort. *Sociology of Education*, *65*(2), 95-113.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(6), 543-553.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*(3-4), 199-215.

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 national achievement tests. *Applied Measurement in Education*, *17*(3), 301-321.

Ercikan, K., Roth, W. M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education*, *27*(4), 273-285.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An Introduction*. CA: Thousand Oaks, Sage Publications.

Geske, A., & Ozola, A. (2010). *Differential item functioning in the aspect of gender differences in reading literacy*. A paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden.

Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *Alberta Journal of Educational Research*, *45*(4), 353-376.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*(2), 164-187.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*(1), 69 -86.

Grossman, F. M., & Franklin, N. K. (1988). Bias effects in speech-language assessment and decision-making. *Language, Speech, and Hearing Services in Schools*, *19*(2), 153-159.

Hernandez, D. J. (2011). Double Jeopardy: How third-grade reading skills and poverty influence high school graduation. *Annie E. Casey Foundation*.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the Military Testing Association*, *1*, 282-287.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel -Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

International Study Center. (2016). TIMSS & PIRLS. Retrieved from http://timssandpirls.bc.edu/about.html.

Jiang, W. (2000). The relationship between culture and language. *ELT journal*, *54*(4), 328 -334.

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, *30*(3), 420-437.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*(4), 437-447.

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology, 45*(3), 381-399.

Knighton, T., & Bussière, P. (2006). *Educational outcomes at age 19 associated with reading ability at age 15*. Ottawa: Statistics Canada.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*(2), 193-220.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation*. Unpublished PhD thesis, University of North Carolina, Chapel Hill, California.

Lau, S. P. (2007). *The history of Macao education*. Macao: Lau Sin Peng. [in Chinese]

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105-118.

Penfield, R. D. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *Alberta Journal of Educational Research*, *49*(3), 231 -243.

Penfield, R. D., & Camilli, G. (2006). 5 Differential item functioning and item bias. *Handbook of Statistics*, *26*, 125-167.

Penfield, R. D. (2012). DIFAS (Version 5.0) [Computer software].

Prieto-Maraņón, P., Aguerri, M. E., Galibert, M. S., & Attorresi, H. F. (2012). Detection of Differential Item Functioning. *Methodology*, *8*(2), 63-70.

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions?. *Applied Measurement in Education*, *24*(4), 349-366.

Organization for Economic Co-operation and Development (2012). *PISA 2009 technical report*. PISA, OECD Publishing, Paris. Retrieved from http://dx.doi.org/10.1787/9789264167872-en

Organization for Economic Co-operation and Development (2010a). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Paris: Author.

Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing*, *13*(2), 152-174.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1992). Evaluating hypotheses about differential item functioning. *ETS Research Report Series*, *1992*(1).

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *25*(1), 1-13.

Shan, P. W. J., & Ieong, S. S.L. (2009). Post-colonial reflections on education development in Macau. A paper presented at the Annual Conference of the Comparative Education Society of Hong Kong, Hong Kong.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147-170.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Soh, K. (2014). Test language effect in international achievement comparisons: An example from PISA 2009. *Cogent Education*, *1*(1), 1-10.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing, 9*(2), 78-91.

Solano-Flores, G., Contreras-Niño, L. Á., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. *Research on PISA*, 71-85.

SPSS. (2016). SPSS (Version 24). [Computer software].

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360-407.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. (2016). Dynamic testing of analogical reasoning in 5-to 6-year-olds: Multiple-choice versus constructed-response training items. *Journal of Psychoeducational Assessment*, *34*(6), 550-565.

Stubbe, T. C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation*, *17*(6), 465-481.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Thissen, D. (2001). IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. *Chapel Hill, NC: LL Thurstone Psychometric Laboratory*.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer & H. Wainer (Eds.), *Differential Item Functioning*, (pp. 67-113). Hillside, NJ: Lawrence Erlbaum.

Thomas, A. H. (1992). Alternatives to retention: If flunking doesn't work, what does?. *OSSC Bulletin*, *35*(6), 6.

Van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

Woods, C. M., Cai, L., & Wang, M. (2012). The Langer-improved Wald test for DIF testing with multiple groups evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547.

Woods, C. M. (2008). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement*, *68*(4), 571-586.

Appendix
PISA 2009 reading passages and questions

MOBILE PHONE SAFETY

Are mobile phones dangerous?

Yes	No
1. Radio waves given off by mobile phones can heat up body tissue, having damaging effects.	Radio waves are not powerful enough to cause heat damage to the body.
2. Magnetic fields created by mobile phones can affect the way that your body cells work.	The magnetic fields are incredibly weak, and so unlikely to affect cells in our body.
3. People who make long mobile phone calls sometimes complain of fatigue, headaches, and loss of concentration.	These effects have never been observed under laboratory conditions and may be due to other factors in modern lifestyles.
4. Mobile phone users are 2.5 times more likely to develop cancer in areas of the brain adjacent to their phone ears.	Researchers admit it’s unclear this increase is linked to using mobile phones.
5. The International Agency for Research on Cancer found a link between childhood Cancer and power lines. Like mobile phones, power lines also emit radiation.	The radiation produced by power lines is a different kind of radiation, with much more energy than that coming from mobile phones.
6. Radio frequency waves similar to those in mobile phones altered the gene expression in nematode worms.	Worms are not humans, so there is no guarantee that our brain cells will react in the same way.

If you use a cell phone…

Do	Don’t
Keep the calls short.	Don’t use your mobile phone when the reception is weak, as the phone needs more power to communicate with the base station, and so the radio-wave emissions are higher.
Carry the mobile phone away from your body when it is on standby.	

Buy a mobile phone with a long “talk time”. It is more efficient, and has less powerful emissions.	Don’t buy a mobile phone with a high “SAR” value[1]. This means that it emits more radiation. Don’t buy protective gadgets unless they have been independently tested.

[1] SAR (specific absorption rate) is a measurement of how much electromagnetic radiation is absorbed by body tissue whilst using a cell phone.

“Mobile Phone Safety” on the previous two pages is from a website.

Use “Mobile Phone Safety” to answer the questions that follow.

Question 1: MOBILE PHONE SAFETY

What is the purpose of the **Key points**?

A. To describe the dangers of using mobile phones.
B. To suggest that debate about mobile phone safety is ongoing.
C. To describe the precautions that people who use mobile phones should take.
D. To suggest that there are no known health problems caused by mobile phones.

Question 2: MOBILE PHONE SAFETY

“It is difficult to prove that one thing has definitely caused another.”

What is the relationship of this piece of information to the Point 4 **Yes** and **No** statements in the table **Are mobile phones dangerous**?

A. It supports the Yes argument but does not prove it.
B. It proves the Yes argument.
C. It supports the No argument but does not prove it.
D. It shows that the No argument is wrong.

Question 3: MOBILE PHONE SAFETY

Look at Point 3 in the No column of the table. In this context, what might one of these “other factors” be? Give a reason for your answer.

Question 4: MOBILE PHONE SAFETY

Look at the table with the heading **If you use a cell phone …**

Which of these ideas is the table based on?

A. There is no danger involved in using mobile phones.
B. There is a proven risk involved in using mobile phones.
C. There may or may not be danger involved in using mobile phones, but it is worth taking precautions.
D. There may or may not be danger involved in using mobile phones, but they should not be used until we know for sure.
E. The **Do** instructions are for those who take the threat seriously, and the **Don't** instructions are for everyone else.

THE PLAY'S THE THING

Take place in a castle by the beach in Italy.

FIRST ACT

Ornate guest room in a very nice beachside castle. Doors on the right and left. Sitting room set in the middle of the stage: couch, table, and two armchairs. Large windows at the back. Starry night. It is dark on the stage. When the curtain goes up we hear men conversing loudly behind the door on the left. The door opens and three tuxedoed gentlemen enter. One turns the light on immediately. They walk to the center in silence and stand around the table. They sit down together, Gál in the armchair to the left, Turai in the one on the right, Ádám on the couch in the middle. Very long, almost awkward silence. Comfortable stretches. Silence Then:

GÁL
Why are you so deep in thought?

TURAI
I'm thinking about how difficult it is to begin a play. To introduce all the principal characters in the beginning, when it all starts.

ÁDÁM
I suppose it must be hard.

GÁL

Quite a peculiar brain you've got. Ca't you forget your profession for a single minute?

TURAI

That cannot be done.

GÁL

Not half an hour passes without you discussing theatre, actors, plays. There are other things in this world.

TURAI
There aren't. I am a dramatist. That is my curse.

GÁL
You shouldn't become such a slave to your profession.

TURAI
If you do not master it, you are its slave. There is no middle ground. Trust me, it's no joke starting a play well. It is one of the toughest problems of stage mechanics. Introducing your characters promptly. Let's look at this scene here, the three of us. Three gentlemen in tuxedoes. Say they enter not this room in this lordly castle, but rather a stage, just when a play begins. They would have to

TURAI

It is – devilishly hard. The play starts. The audience goes quiet. The actors enter the stage and the torment begins. It's an eternity, sometimes as much as a quarter of an hour before the audience finds out who's who and what they are all up to.

GÁL

Stands up. My name is Gál, I'm also a playwright. I write plays as well, all of them in the company of this gentleman here. We are a famous playwright duo. All playbills of good comedies and operettas read: written by Gál and Turai. Naturally, this is my profession as well.

GÁL and TURAI

Together. And this young man…

ÁDÁM

Stands up. This young man is, if you allow me, Albert Ádám, twenty-five years old, composer I wrote the music for these kind gentlemen for their latest operetta. This is my first work for the stage. These two elderly angels have discovered me and now, with their help, I'd like to become famous. They got me invited to this castle. They got my dress-coat and tuxedo made. In other words, I am poor and unknown, for now. Other than that I'm an orphan and my grandmother raised me. My grandmother has passed away. I am all alone in this world. I have no name, I have no money.

TURAI

But you are young.

GÁL

And gifted.

ÁDÁM

And I am in love with the soloist.

chat about a whole lot of uninteresting topics until it came out who we are. Wouldn't it be much easier to start all this by standing up and introducing ourselves? *Stands up.* Good evening. The three of us are guests in this castle. We have just arrived from the dining room where we had an excellent dinner and drank two bottles of champagne. My name is Sándor Turai, I'm a playwright, I've been writing plays for thirty years, that's my profession. Full stop. Your turn.

TURAI

Now wouldn't this be the easiest way to start a play?

GÁL

If we were allowed to do this, it would be easy to write plays.

TURAI

Trust me, it's not that hard. Just think of this whole thing as …

GÁL

All right, all right, all right, just don't start talking about the theatre again. I'm fed up with it. We'll talk tomorrow, If you wish.

TURAI You shouldn’t have added that. Everyone in the audience would figure that out anyway. *They all sit down*.	

“The Play’s the Thing” is the beginning of a play by the Hungarian dramatist Ferenc Molnár.

Use “The Play’s the Thing” on the previous two pages to answer the questions that follow. (Note that line numbers are given in the margin of the script to help you find parts that are referred to in the questions.)

Question 1: THE PLAY’S THE THING

What were the characters in the play doing **just before** the curtain went up?

Question 2: THE PLAY’S THE THING

“It’s an eternity, sometimes as much as a quarter of an hour.” (lines 29-30)

According to Turai, why is a quarter of an hour “an eternity”?

A. It is a long time to expect an audience to sit still in a crowded theatre.
B. It seems to take forever for the situation to be clarified at the beginning of a play.
C. It always seems to take a long time for a dramatist to write the beginning of a play.
D. It seems that time moves slowly when a significant event is happening in a play.

Question 3: THE PLAY’S THE THING

A reader said, “Ádám is probably the most excited of the three characters about staying at the castle.”

What could the reader say to support this opinion? Use the text to give a reason for your answer.

Question 4: THE PLAY’S THE THING

Overall, what is the dramatist Molnár doing in this extract?

A. He is showing the way that each character will solve his own problems.
B. He is making his characters demonstrate what an eternity in a play is like.
C. He is giving an example of a typical and traditional opening scene for a play.
D. He is using the characters to act out one of his own creative problems.

TELECOMMUTING

The way of the future

Just imagine how wonderful it would be to "telecommute"[1] to work on the electronic highway, with all your work done on a computer or by phone! No longer would you have to jam your body into crowded uses or trains or waste hours and hours travelling to and from work. You could work wherever you want to – just think of all the job opportunities this would open up!

Molly

Disaster in the making

Cutting down on commuting hours and reducing the energy consumption involved is obviously a good idea. But such a goal should be accomplished by improving public transportation or by ensuring that workplaces are located near where people live. The ambitious idea that telecommuting should be part of everyone's way of life will only lead people to become more and more self-absorbed. Do we really want our sense of being part of a community to deteriorate even further?

Richard

[1] "Telecommuting" is a term coined by Jack Nilles in the early 1970s to describe a situation in which workers work on a computer away from a central office (for example, at home) and transmit data and documents to the central office via telephone lines.

Use "Telecommuting" above to answer the questions that follow.

Question 1: TELECOMMUTING

What is the relationship between "The way of the future" and "Disaster in the making"?

A. They use different arguments to reach the same general conclusion.
B. They are written in the same style but they are about completely different topics.
C. They express the same general point of view, but arrive at different conclusions.
D. They express opposing points of view on the same topic.

Question 2: TELECOMMUTING

What is one kind of work for which it would be difficult to telecommute? Give a reason for your answer.

Question 3: TELECOMMUTING

Which statement would **both** Molly and Richard agree with?

A. People should be allowed to work for as many hours as they want to.
B. It is not a good idea for people to spend too much time getting to work.
C. Telecommuting would not work for everyone.
D. Forming social relationships is the most important part of work.