Examining test fairness across gender in a computerised reading test: A comparison between the Rasch-based DIF technique and MIMIC

Xuelian Zhu[1] & Vahid Aryadoust[2]
[1] Sichuan International Studies University, China; National Institute of Education, Nanyang Technological University, Singapore
[2] National Institute of Education, Nanyang Technological University, Singapore

Test fairness has been recognised as a fundamental requirement of test validation. Two quantitative approaches to investigate test fairness, the Rasch-based differential item functioning (DIF) detection method and a measurement invariance technique called multiple indicators, multiple causes (MIMIC), were adopted and compared in a test fairness study of the Pearson Test of English (PTE) Academic Reading test ($n = 783$). The Rasch partial credit model (PCM) showed no statistically significant uniform DIF across gender and, similarly, the MIMIC analysis showed that measurement invariance was maintained in the test. However, six pairs of significant non-uniform DIF ($p < 0.05$) were found in the DIF analysis. A discussion of the results and post-hoc content analysis is presented and the theoretical and practical implications of the study for test developers and language assessment are discussed.

Key words: Pearson Test of English (PTE) Academic Reading test, Rasch-based partial credit model, MIMIC, differential item functioning (DIF), measurement invariance

Introduction

Test fairness, a crucial concept in language testing and assessment, is broadly defined as the equitable treatment of test takers, the lack of bias in measurement, and justifiable uses and interpretations of test scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Although different conceptual frameworks of test fairness have incorporated broader concepts such as the validity or the socioeconomic aspects of tests in test fairness (McNamara, Knoch, & Fan, 2019; Xi, 2010), in a narrow

Email address for correspondence: vahid.aryadoust@nie.edu.sg

sense, fairness is concerned with the consistency of test functions, most notably whether the background characteristics of test takers affect their performance on language tests (McNamara & Ryan, 2011). We adopt McNamara and Ryan's (2011) definition of test fairness in the present study, which refers to "the extent to which the test quality, especially its psychometric quality, ensures procedural equality for individual and subgroups of test-takers and the adequacy of the representation of the construct in test materials and procedures" (p. 163). To investigate test fairness in this sense, two statistical procedures have been adopted by researchers: differential item functioning (DIF) investigation across sub-groups at item level (Ferne & Rupp, 2007; Zumbo, Liu, Wu, Shear, Olvera Astivia & Ark, 2015) and measurement invariance (MI) of the factor structure of tests across groups (Ginther & Stevens, 1998; Stricker, Rock & Lee, 2005; Youn & Im, 2016). Due to the differences in statistical rationales behind the two techniques, there is a possibility that they yield different results and therefore have different implications for test fairness (Borsboom, Mellenbergh, & van Heerden, 2002). Comparing different techniques in the context of high-stakes tests could help to make more informed decisions concerning test fairness by taking into account the differences between the techniques and the implications they have for test fairness.

DIF refers to the extent to which test items function differently for different subgroups of test takers with the same ability level because the test is measuring off-trait characteristics such as gender, mother tongue or academic background, etc. (Banerjee & Papageorgiou, 2016). DIF analysis has been used to provide evidence as to whether the test actually measures the ability intended to be measured and does not put any groups in a favorable or unfavorable position, thus shedding light on the sources of construct-irrelevant variance in test scores (Messick, 1989). Common methods to detect DIF include the Mantel-Haenszel (MH) method which requires discretization of the conditioning variable and assumes no interaction between test takers' attributes and items (Gallagher, 2004; Holland & Thayer, 1988; Mantel & Haenszel, 1959), the logistic regression (LogR) models which allow for an interaction between test takers' attributes and items (Swaminathan & Rogers, 1990), and item response theory (IRT) models which estimate item difficulty and test takers' individual ability within a single probabilistic framework (Kim, 2001; Muraki, 1999; Pae, 2004; Rasch, 1960; Wright & Stone, 1979). Among these techniques, Rasch-based DIF analysis, a commonly used method of validation (McNamara & Knoch, 2012), offers the possibility of detecting both uniform DIF and non-uniform DIF (see Rasch-based DIF analysis). The Rasch method computes item difficulty across groups without being influenced by other parameters such as guessing or discrimination, which renders it more convenient and less demanding in terms of the sample size than other item response theory (IRT) models (Rouquette, Hardouin, Vanhaesebrouck, Sébille, & Coste, 2019).

The second method to investigate test fairness is based on measurement invariance (MI), which is known as factorial invariance or measurement equivalence. MI techniques are used to investigate whether tests measure the latent trait under investigation in the same way across groups (Liu & Dorans, 2016; Yoo, Manna, Monfils, & Oh, 2018). Compared with DIF analysis, which is used to evaluate test fairness by calculating the number of DIF items, MI is defined as a bipartite situation wherein there is no statistical difference between different parameters such as regression weights and error terms across different groups (Borsboom, 2006). Confirmatory factor analysis (CFA), which is a special case of structural equation modeling (SEM), is commonly employed to assess MI by looking into factor loading coefficients, item intercepts/thresholds, residual variances, and covariances. One commonly used CFA-based method of investigating MI is multiple indicators, multiple causes (MIMIC) that functions in a similar way as DIF (Kim, Yoon & Lee, 2012). Although both DIF and MIMIC are used to compare subgroups of test takers to validate tests and avoid bias, there is a dearth of research in comparing their potential in understanding fairness in language assessment. In the present study, these methods are adopted and compared to examine fairness in the Pearson Test of English (PTE) Academic Reading test, a high-stakes test of English proficiency.

The PTE Academic was developed and operationalised in 2009. It is administered through computers and its test scores are used by governments, employers and university officials to decide whether immigration, job, or university applicants have achieved certain language proficiency to study and/or work in English speaking environments (Pearson, n.d.). Since the PTE Academic acts as a gatekeeper to higher learning, employment or immigration, it would be crucial for its designers to demonstrate that it is a fair measure of the English language proficiency of test takers from all backgrounds (Riazi, 2013; Wang, Choi, Schmidgall, & Bachman, 2012). Despite the significant role of the PTE Academic in test candidates' academic life, however, there is scant research on the evidential basis of its test fairness, which is likely due to its fairly recent introduction to the language assessment market and its unique format as a computer-assisted language assessment (Song, 2014; H. K. Pae, 2012; Zheng & De Jong, 2011). Under such a backdrop, we conducted a Rasch-based DIF analysis and a MIMIC analysis on the PTE Academic Reading test to provide stakeholders with evidence of test fairness and offer methodological guidance to the language assessment community.

Literature Review

The Rasch-based DIF Analysis

DIF exists when different groups of test takers of the same ability level have significantly different chances of answering a test item because the test interacts with off-trait characteristics such as test takers' gender or mother tongue (Engelhard, 2012). DIF analysis has been applied to investigate the validity of the uses and interpretations of test scores, as the presence of DIF might indicate the undesirable effect of some factors that are not relevant to the construct under assessment. The presence of DIF has also been regarded as a violation or attenuation of test fairness by McNamara and Ryan (2011) who alluded to Messick's (1989) validity framework to argue that investigating fairness would provide the evidential basis for the better utilization and interpretation of test scores.

From a psychometric perspective, there are a number of factors which can adversely affect test fairness and which the designers of high-stakes tests should monitor closely such as gender (Chubbuck, Curley & King, 2016; Ryan & Bachman, 1992), age (Geranpayeh & Kunnan, 2007), language background (Harding, 2011), ethnicity (Oliveri, Lawless, Robin & Bridgeman, 2018; Zeidner, 1987), and academic experience (Alderson & Urquhart, 1985; Pae, 2004). Research has shown that these factors can yield significant DIF and put certain groups to an advantage or disadvantage. To exert sizeable impacts on test takers' measured performances, the detected DIF must be statistically significant ($p < .05$) and substantive (Bond & Fox, 2015; Linacre, 2018a); in this case, DIF would indicate that test scores are confounded by unintended factors beyond the construct which jeopardise test fairness (Messick, 1996; Wright & Stone, 1999).

There are multiple DIF analysis techniques, such as the Rasch models (Raquel, 2019; Rasch, 1960; Wright & Stone, 1979), the Rasch trees (e.g., Aryadoust, 2018; Strobl, Kopf & Zeileis, 2015), multidimensional item response theory models (Ackerman, 1992; Reckase, 1997), the Mantel-Haenszel method (Holland & Thayer, 1988), the logistic regression method (LR; Swaminathan & Rogers, 1990), the standardization procedure (Dorans & Kulick, 1983, 1986), and the simultaneous item bias test (SIBTEST) (Shealy & Stout, 1993). Many of these techniques, such as Rasch measurement, are based on the precondition that person measures should remain invariant (within error) across different measurement conditions (Bond & Fox, 2015). Otherwise, failure to establish invariance across test takers' background results in DIF and the effect of construct-irrelevant factors on test scores (Aryadoust, Goh & Lee, 2011; Pae, 2004; Takala & Kaftandjieva, 2000). Rasch-based DIF analysis first computes the local difficulty parameter with a standard error of measurement for each group and compares the differences of these difficulty parameters to generate a DIF contrast. Next, a two-sided t-test comparing the difference between the two means is conducted to produce the DIF significance as a Welch t value. The null hypothesis in DIF analysis is that the two means are the same within the range of measurement error. This hypothesis is rejected when the DIF contrast of these two means is statistically significant.

SEM-based MIMIC

Structural equation modeling (SEM) has been applied in language testing and assessment as a confirmatory approach to model testing (In'nami & Koizumi, 2011; Kline, 2015; Ockey & Choi, 2015). SEM is used to anlyse the relationships among observed and latent variables, conferring advantages over the traditional path analysis which only models observed variables (Byrne, 2016). Moreover, SEM can analyse multivariate relationships or indirect effects, which is difficult to achieve by other statistical tools (In'nami & Koizumi, 2011). Therefore, it is possible to use SEM to determine whether an independent variable could directly affect a dependent variable or through another variable regardless of whether they are latent or observed.

MIMIC is a method to examine test invariance across two or multiple groups by regressing latent variables and indicators (items) on covariates representing subpopulations (e.g., gender) (Jöreskog & Goldberger, 1975). There are two basic procedures in MIMIC modeling: (1) establish a full sample model; (2) add the covariates to investigate their effects on the latent variables and indicators (Brown, 2015). If the covariates under investigation elicit no effects on individual test items, there would be evidence supporting the invariance of measurement (Lúcio et al., 2017). Otherwise, the results could be regarded as evidence of measurement non-invariance, which would indicate that the affected test item(s) would influence the score interpretation of a group of candidates beyond their measured (latent) ability. Since MIMIC incorporates grouping variables like gender as covariates instead of testing the model in each group, it can significantly reduce error in the analysis procedure, thus helping to facilitate the analysis of MI across different backgrounds.

DIF across Gender

In language assessment, even though gender is regarded as one of the "more stable" test taker characteristics, which means that it is not likely to elicit differential treatment on test takers compared to the "less stable" characteristics such as strategies, skills, and motivations (Alderson, 2000, p. 56), a number of studies have reported DIF across female and male test takers (Carlton & Harris, 1992; Lawrence, Curly & McHale, 1988; Ryan & Bachman, 1992; Pae, 2004). This suggests that the constructs under investigation were measured in different ways for females and males even if they match on the overall aptitude scores (i.e., ability level) (Camilli & Shepard, 1994). Further, researchers have attempted to explain that the content and features of reading passages, item types, or cognitive skills of test takers could contribute to the differential performance of gender groups on tests (T. I. Pae, 2012).

Geranpayeh and Kunnan (2007) synthesised empirical studies in language testing which focused on differentiated gender performance, such as Zeidner (1986, 1987), Ryan and Bachman (1992), and Takala and Kaftandjieva (2000). Among these, Ryan and Bachman (1992) found that four out of 140 items in Test of English as a Foreign Language (TOEFL) favored males whereas two items favored females, which accounted for a small portion of the items and, therefore, their influence on the entire test would be negligible. Conversely, Takala and Kaftandjieva (2000), who investigated the Finnish Foreign Language Certificate Examination (FFLCE), found 11 items confounded by gender DIF of which six items favored males and five favored females, suggesting that the DIF items would cancel each other out in five cases, leaving only one DIF item functioning in favor of males. Here, although the effects of bias appear at the item level, the total test score is not necessarily biased, since, based on the cancellation rule in DIF, items advantaging and disadvantaging different groups of test takers cancel each other out (Borsboom, 2006; Teresi, 2006). That is why the appearance of DIF is "a necessary but not sufficient condition" for the violation of test fairness (McNamara & Roever, 2006, p. 83). For DIF to be considered bias or the violation of test fairness, the majority or all items should systematically advantage one specific group of test takers.

Van Langen, Boskers and Dekkers (2006) studied test takers' performance on the Programme for International Student Assessment (PISA) and reported that reading literacy interacted highly with gender and country and that the differentiation between gender was likely due to some cognitive differences. Park (2008) studied gender differences in the Korea College Scholastic Ability Test (KCSAT), identifying 6 out of 17 items favoring males, while 7 favoring females. Through further content analysis, Park claimed that most items with pictures favored males while most items asking for factual information were easier for females. In another study, Aryadoust, Goh and Lee (2011) found gender DIF in the Michigan English Language Assessment Battery (MELAB) listening test and explained that the lower ability males were likely to use lucky guesses successfully. T. I. Pae (2012) investigated the effect of gender DIF in the KCSAT items, and identified reading strategies and the perceived interest as two possible reasons for the presence of gender DIF. For example, females were favored by the items measuring mood, impressions or tones whereas males were favored by the items in the fill-in-the-blank format. H. I. Pae, (2012), in a separate study, conducted a small-scale DIF analysis across 140 test takers on the PTE Academic field test, finding no significant DIF across gender, while he pointed out that since the study was built on just one form of field tests, real-exam data should be used to validate the test scores.

Further studies have recently examined DIF in test content and forms. For example, Chubbuck, Curley, and King (2016) found that most sports and science materials in the critical reading sections of the Scholastic Assessment Test (SAT) were DIF-free across gender through the MH statistic method, contradicting with previous findings which had claimed these materials would advantage males (SAT Update, 2011). Meanwhile, they suggested that using more than one type of statistical method to investigate DIF could reduce the chances of generating misleading results. Through further interviews, the researchers noted that the familiarity of the topic had little effect on test performance, but the students' overall ability mattered. In another study, Wedman (2017), who analysed test data from five administrations of the Swedish Scholastic Assessment Test (SweSAT), found that, overall, males outperformed females in the verbal section of the test that includes vocabulary, reading comprehension, and sentence completion. However, a closer look revealed that vocabulary items which were selected from the female domain favored females, but items sampled from the male domain did not necessarily favor males, while sentence completion items in reading comprehension favored male participants.

Although much can be learned from the previous studies, it is essential to conduct gender DIF analysis on the PTE Academic Reading test. First, the gender-related DIF research has frequently shown that gender is a noticeable factor influencing fairness in large-scale high-stakes tests (Zumbo 2007). Accordingly, it is necessary to conduct DIF analysis in different test situations to help test designers to optimise or delete gender-sensitive items and provide better estimates of true abilities of test takers. Second, the reasons for gender-based DIF may derive from sources of construct-irrelevant variance such as lucky guesses made by low ability male test takers (Aryadoust et al., 2011), but these sources require evidence from more empirical studies to confirm. Moreover, even if gender-related DIF is detected by certain statistical procedures, other statistical tools could be adopted to compare the results, thus enhancing the reliability of DIF investigations (Suh & Talley, 2015; Teresi, 2006). To address these gaps, two research questions are investigated in this study.

1. Does the PTE Academic Reading test display gender-related DIF? Is DIF mediated by the ability level of test takers?

2. Do the PTE Academic Reading test and its items function equivalently between different gender groups?

Methodology

Participants

Participants of this study were 783 test takers aged 17 to 39 (M = 27.2; SD = 5.28), 281 (35.9%) of whom are females and 502 (64.1%) are males. Other demographic information includes participants' home language spoken as mother tongue, country of birth, and country of citizenship that identifies their official nationality. Among them, 472 (59.0%) test takers are from India, and the next three largest groups of test takers by country of origin are those from Nepal (64, 8.0%), Pakistan (52, 6.6%), and China (41, 5.0%). The home languages spoken by the majority of test takers are Punjabi

(127, 16.2%), Telugu (100, 12.8%), English (110, 14.0%) and Hindi (91, 11.6%). The participants took PTE Academic test between January 2015 and March 2016, and their performances including their total scores in the reading test and subscores on 10 reading items were recorded via computers at secured test centers.

Instruments

The PTE Academic test has been conducted in secured test centers around the world and scored on computers since its first implementation. The Reading test prompt used in this study included 10 items, among which eight are using partial credit marking schemes and two are dichotomous. Table 1 is a summary of the test prompt. For example, Item 1 is a banked choice item providing a pool of 25 choices for the test takers to choose from. Among the available choices, only five choices would fit the given blanks, each carrying one point. This makes the test item a partial credit question with six possible categories of scoring: 0, 1, 2, 3, 4, and 5.

Table 1. The PTE Academic Reading Test

Item	Scoring	Item Type
1	Partial credit (6 categories)	Banked choices: 5/25 choices
2	Partial credit (7 categories)	Banked choices: 6/24 choices
3	Partial credit (3 categories)	MCQ: 2/5 options
4	Partial credit (3 categories)	MCQ: 2/5 options
5	Partial credit (4 categories)	Reordering: 4 options
6	Partial credit (5 categories)	Reordering: 5 options
7	Partial credit (6 categories)	Banked choices: 5/8 choices
8	Partial credit (5 categories)	Banked choices: 4/7 choices
9	Dichotomous (2 categories)	MCQ: 1/4 options
10	Dichotomous (2 categories)	MCQ: 1/4 options

Data Analysis

Two rounds of data analysis were performed to answer the two research questions concerning the existence of DIF. The first round is Rasch-based DIF analysis and the second round is SEM-based MIMIC model analysis.

Rasch-based DIF analysis

A pairwise t-test was used in the Rasch-based analysis. The test items in the present study have different scoring categories and, accordingly, the partial credit model (PCM) was applied (Wright & Masters, 1982). PCM provides various fit statistics comprising infit and outfit mean square (MnSq) and Standardised Z values (Z-STD) to determine whether the data fit the model. Infit, an information-weighted sum, is sensitive to unexpected response patterns when persons perform on items with nearly the same level of difficulty, and outfit, based on the conventional sum of squared standardised residuals, is sensitive to unexpected answers to very difficult or very easy items (Bond & Fox, 2015; Linacre, 2018b). Other statistics generated through PCM analysis included reliability and separation which were used to measure the replicability of person or item locations along the latent trait continuum. The higher reliability and separation are, the more accurate the order of the item or person within the sample. In addition, we performed a principal component of residuals (PCAR) to identify whether substantive secondary dimensions or contrasts (components that explain the largest possible amount of variance in the residuals) exist in Rasch residuals. Contrasts with eigenvalue less than 2 do not have the power to create a meaningful influence on the construct under assessment (Linacre, 2018b). To examine local independence, which indicates whether a test taker's performance on an item is influenced by other items, the residual correlations or Q3 statistics were computed, and correlations need to be around 0.7 or more to indicate dependency among items (Linacre, 2018b). The Winsteps computer package, Version 4.4, (Linacre, 2018a) was employed to run the Rasch-based PCM analysis to investigate gender DIF.

Uniform and non-uniform DIF analyses were performed. Uniform DIF (UDIF) occurs when DIF is consistent across all levels of ability in the subgroup, whereas non-uniform DIF (NUDIF) holds when DIF varies with ability level. In both analyses, significance and substantiality are meaningful. If the p value of the observed DIF is below .05, the next step is to check whether the DIF contrast is large enough to be considered substantive. The power of significant DIF ($p < 0.01$) can be classified into three categories based on the DIF contrast indices: A-level (< 0.43 logits) indicating negligible DIF or very low power; B level (0.43 logits - 0.64 logits) indicating slight to moderate DIF; and C level (≥ 0.64 logits) indicating moderate to large DIF or high power (Linacre, 2018b; Zwick, Thayer, Lewis, 1999).

SEM-based MIMIC analysis

Next, measurement invariance was tested, using a series of SEM-based MIMIC models. The models were estimated with IBM SPSS AMOS, Version 24.0 (IBM Corp, 2016). AMOS is suitable for analyzing both polytomous and dichotomous data (Arbuckle, 2006). First, we investigated the model fit and, then, we statistically regressed the endogenous variables (dependent variables consisting of the latent variable and the test items) on the exogenous variable (independent variable, which was gender). We used the Maximum Likelihood (ML) method of parameter estimation to estimate the fit of the MIMIC models. The model fit was evaluated, using (1) the $\chi 2$ test, (2) Normed Fit Index (NFI) (3) Comparative Fit Index (CFI); (4) Tucker-Lewis Fit Index (TLI), and (5) Root Mean Square Error of Approximation (RMSEA).

Figure 1 presents the components of the SEM-based MIMIC for Item 1 of the test under investigation. The first component of the model is a confirmatory factor analysis (CFA) model with one latent variable (Reading Ability) which is the reading ability measured by the reading test, and its 10 observed variables (10 items). There are 10 single-headed arrows running from Reading Ability to 10 items individually. The other part is a structural model consisting of the observed variable gender and one latent variable (Reading Ability) to establish a predictive relationship between these two variables.

In this study, gender was coded as 1 for females, and 2 for males. Positive regression coefficients for gender-item regression paths would, therefore, indicate that males would have a higher chance to score highly on the items than females. Standardised and unstandardised regression weights of the gender-item paths were calculated in the MIMIC models. Standardised regression weights are normalised coefficients that do not have units and are useful to compare effects across different measures, whereas the unstandardised are in original units, represent the relationship between the raw data, and are useful when variables in the equation use the same scale.

In addition, there are also one-headed arrows from gender to Items 1 through 10 in Figure 1. The blue-black color code indicates that items were regressed on the latent variable one at a time, generating 10 separate MIMIC models each with the aforementioned fit statistics. The relationship between gender and the latent variable fixes (conditions) the ability level of the test takers in the analysis and then through computing the gender-item relationships we estimated the size of DIF (Byrne, 2016). According to Byrne (2016), if the relationship between gender and the item under investigation is significant (p < .05), then the item has violated MI.

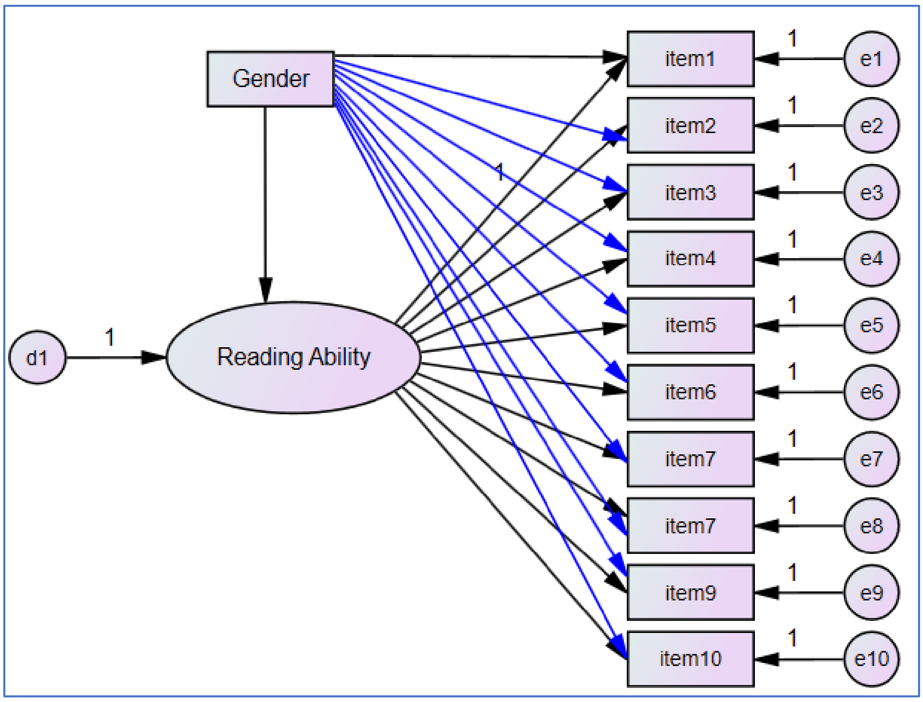

Figure 1. MIMIC model for item 1 to 10

Results

Rasch-based DIF Analysis

Prior to the DIF analysis, the dimensionality check through PCAR showed the amount of raw variance explained by measures was 47.4% (eigenvalues = 9.00), of which 16.9% (eigenvalues = 3.20) was explained by person measures and 30.5% (eigenvalues = 5.80) by items. The raw unexplained variance was 52.6% (eigenvalues = 10.00) of the total raw variance in observations and the first contrast (dimension) in the residuals explained 8.8% of the total variance (eigenvalues = 1.7). The eigenvalue of 1.7 (< 2.0) indicates that there is no substantive secondary construct beside the primary construct under assessment. In addition, the raw variance explained by measures was approximately 5.29 times (9/1.7 eigenvalues) larger than that in the first contrast. The PCAR analysis, therefore, lends support to the assumption that the data are unidimensional.

When it comes to local independence, we found that the correlation indices were smaller than 0.3 (-0.28 to +0.12), suggesting that the items were independent from each other, did not include shared dimensions beyond the construct, and did not replicate measurement features of each other.

Table 2 reports gender UDIF investigation results. DIF contrast is the difference between the local difficulty measures for females and males (Linacre, 2018b). For example, the local difficulty measure of Item 1 for the female group is -0.42 logits with the standard error of measurement of 0.06 logits, which means the actual difficulty of the item for the female group falls between -0.42 ± 0.06 logits. Similarly, the local difficulty of the item for the male group is -0.45 ± 0.04 logits. DIF contrast between female group and male group is 0.03 logits ([-0.42] – [-0.45]), indicating that Item 1 is 0.03 logits more difficult for the female group compared to the male group. The Welch t value of this contrast is 0.37 and the p value is 0.7114, which is not significant (p < .05). Through the overall observation of the Welch p values from item 1 to 10, it can be seen that UDIF is neither statistically significant nor substantive for the items.

Table 2. Results of UDIF Investigation for Gender

Item	Class A	Measure	DIF SE	Class B	Measure	DIF SE	DIF Contrast	Rasch-Welch	
								t	Welch *p*
1	Female	-0.42	0.06	Male	-0.45	0.04	0.03	0.37	0.7114
2	Female	-0.18	0.06	Male	-0.21	0.04	0.03	0.39	0.6963
3	Female	0.16	0.08	Male	0.06	0.06	0.10	1.03	0.3032
4	Female	0.49	0.09	Male	0.56	0.07	-0.07	-0.64	0.5195
5	Female	-1.42	0.10	Male	-1.33	0.07	-0.09	-0.75	0.4554
6	Female	-0.98	0.07	Male	-0.91	0.05	-0.06	-0.75	0.4523
7	Female	-0.31	0.05	Male	-0.33	0.04	0.02	0.33	0.7426
8	Female	0.53	0.06	Male	0.53	0.05	0.00	0.00	1.0000

9	Female	0.65	0.14	Male	0.95	0.11	-0.30	-1.73	0.0847
10	Female	1.26	0.15	Male	1.28	0.12	-0.20	-0.11	0.9154

Notes: *N* = 783; Female (*n* = 281); Male (*n* = 502)

NUDIF Investigation Results

Table 3 presents NUDIF results across gender. Both the male and female groups were divided into two subgroups based on their performance: high ability female (High F), high ability male (High M), low ability female (Low F), low ability male (Low M). Winsteps invoked 40 NUDIF comparisons (2 female ability subgroups × 2 male ability subgroups × 10 items). This analysis identified items 1, 4, 5, and 9 as having statistically significant NUDIF ($p < .05$).

Figure 2 presents the item characteristic curves (ICCs) of the NUDIF items. The horizontal axis represents item difficulty and the vertical axis shows the test takers' performance. The solid line is the Rasch model trajectory and the lines with different colors represent the groups. The interaction of ICC curves in Item 1, 4, 5 and 9 indicates that the subgroups' probabilities of answering the items correctly varied with ability levels. The specific NUDIF numbers and details are shown in Table 3.

Overall, there were six confirmed NUDIF interactions, which are shown in Table 3. With regard to the level of DIF, the UNDIF contrast between Low F and High F groups on Item 1 (NO.1), High F and High M on Item 1 (NO.2) and Low F and High F on Item 4 (NO.3) were 0.33 logits, 0.42 logits, and 0.41 logits, respectively, which indicates they were A-level (negligible) NUDIF. The DIF contrast between Low F and High M on Item 4 (NO.4), on the other hand, was 0.50 logits, indicating a B-level (slight to moderate) NUDIF. The contrast between Low F and High M on Item 5 (NO.5), and Low F and High F on Item 9 (NO.6) was 0.73 logits and 1.07 logits, suggesting two C-level (moderate to large) NUDIF.

In addition, Item 1 was 0.33 logits more difficult for Low F ($p = 0.0254$) and 0.42 logits more difficult for High F ($p = 0.0331$). By contrast, Low F was likely to perform better on Item 4 compared to either High F and High M, which were statistically significant NUDIF ($p = 0.0255$; $p = 0.0019$). Item 5 caused a significant NUDIF ($p = 0.0006$), favoring the Low F over High M, while Item 9 favored High F over Low F by about 1.07 logits with a NUDIF ($p = 0.0021$).

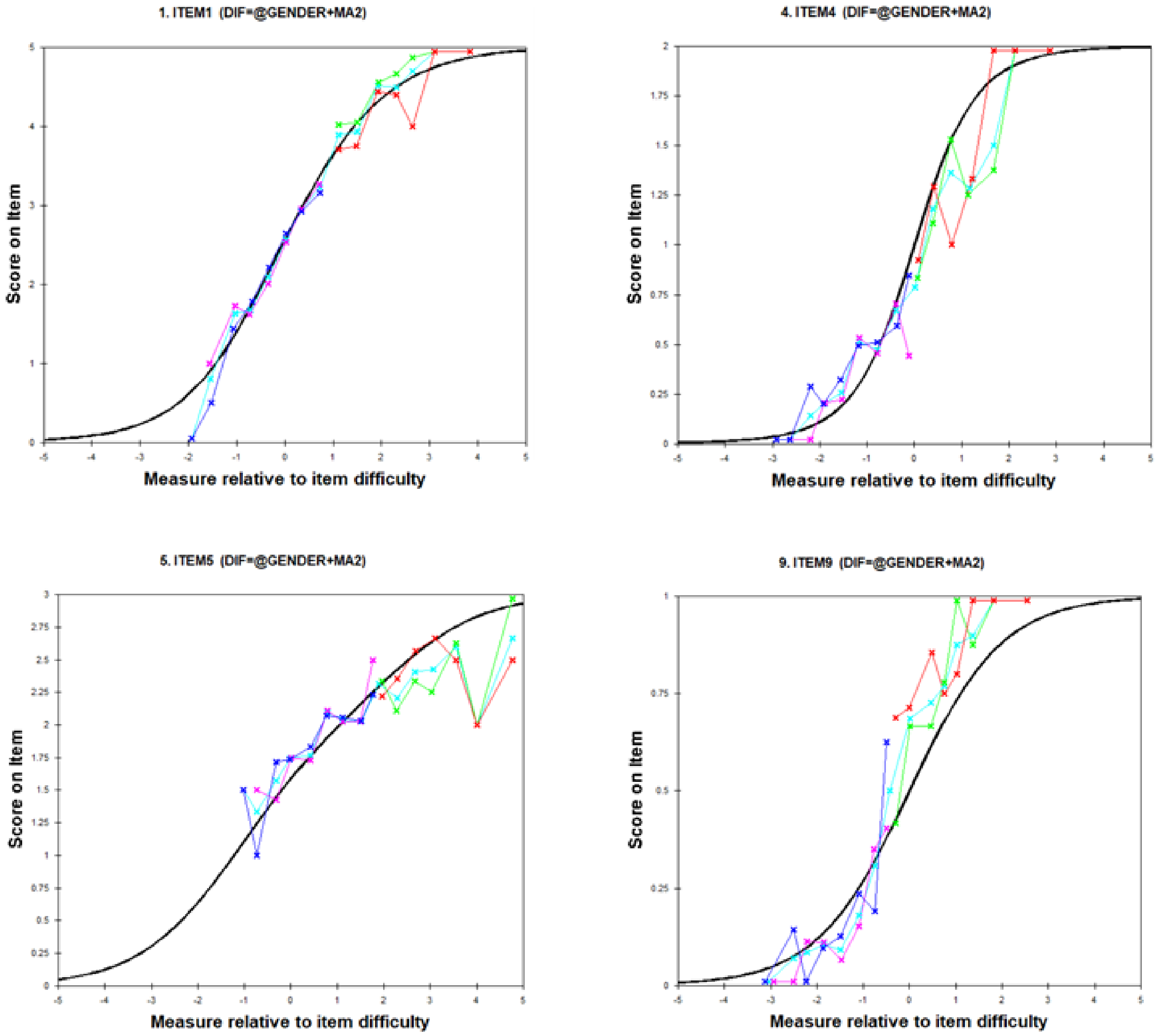

Figure 2. Item characteristic curves of Item 1, 4, 5 and 9

Table 3. Gender NUDIF Analysis Results

DIF number (NO.)	Item	Class A	Measure	DIF SE	Class B	Measure	DIF SE	DIF Contrast	Rasch-Welch			Level of NUDIF
									t	d.f.	Welch *p*	
1	1	Low F	-0.45	0.06	High M	-0.78	0.13	0.33	2.26	150	0.0254	A-level
2	1	High F	-0.36	0.14	High M	-0.78	0.13	0.42	2.15	150	0.0331	A-level
3	4	Low F	0.37	0.10	High F	0.78	0.15	-0.41	-2.26	123	0.0255	A-level
4	4	Low F	0.37	0.10	High M	0.86	0.12	-0.50	-3.14	225	0.0019	B-level
5	5	Low F	-1.52	0.11	High M	-0.79	0.17	-0.73	-3.50	184	0.0006	C-level
6	9	Low F	0.95	0.17	High F	-0.13	0.29	1.07	3.14	114	0.0021	C-level

Notes: N = 783;

Low F = Low-ability female subgroup (n = 215); High F = High-ability female subgroup (n = 66);

Low M = Low-ability male subgroup (n = 401); High M = High-ability male subgroup (n = 101).

Results of the SEM-based MIMIC Analysis

Before the MIMIC analysis, a univariate normality check was run first by calculating the skewness and kurtosis of each item since the maximum likelihood method of parameter estimation adopted in the SEM analysis of this research assumes data normality. In addition, multivariate normality was examined using Mardia's coefficient (Kline, 2011). Table 4 displays an even distribution of skewness and kurtosis indices between -1.96 and +1.96, supporting univariate normality (Field, 2018). In addition, Mardia's coefficient is -4.55, which supports multivariate normality (Kline, 2011).

Table 4. Item-level descriptive statistics

Item	Min.	Max.	Mean	SD	Skewness	Kurtosis
1	0.00	5.00	2.82	1.33	-0.18	-0.70
2	0.00	6.00	3.14	1.39	-0.09	-0.45
3	0.00	2.00	0.83	0.90	0.34	-1.67
4	0.00	2.00	0.61	0.82	0.83	-1.02
5	0.00	3.00	2.04	0.67	-0.32	0.11
6	0.00	4.00	2.62	1.08	-0.41	-0.56
7	0.00	5.00	2.66	1.52	-0.05	-0.91
8	0.00	4.00	1.33	1.25	0.58	-0.72
9	0.00	1.00	0.29	0.46	0.91	-1.17
10	0.00	1.00	0.22	0.42	1.35	-0.18

The model fit indices of the MIMIC analyses are displayed in Table 5. The main model is the MIMIC model in which the items were not regressed on the latent variable in a one by one manner. Likely due to the sample size (n = 783), which is fairly large, the chi-square's p-value reached statistical significance ($p < .05$). However, the fit statistics showed good model-to-data fit in the main model without gender included ($\chi 2$ = 160.684; NFI= 0.875; CFI = 0.905; TLI = 0.881; RMSEA = 0.058 [LO 90 = 0.049; HI 90 = 0.068]) and the model fit for Item 1 to Item 10 also demonstrates good fit. LO 90 and HI 90 are the lower and upper limits of a 90% confidence interval for RMSEA.

Table 5. Fit Statistics of the MIMIC Models

Model	χ^2	DF	p value	NFI	TLI	CFI	RMSEA	LO 90	HI 90
Main model	160.684	44	0.000	0.875	0.881	0.905	0.058	0.049	0.068
Item 1	160.141	43	0.000	0.875	0.878	0.905	0.059	0.049	0.069
Item 2	160.091	43	0.000	0.875	0.878	0.905	0.059	0.049	0.069
Item 3	159.594	43	0.000	0.876	0.879	0.905	0.059	0.049	0.069
Item 4	160.166	43	0.000	0.875	0.878	0.905	0.059	0.049	0.069
Item 5	160.245	43	0.000	0.875	0.878	0.905	0.059	0.049	0.069
Item 6	160.399	43	0.000	0.875	0.878	0.905	0.059	0.049	0.069
Item 7	160.056	43	0.000	0.875	0.878	0.905	0.059	0.049	0.069

Item 8	160.598	43	0.000	0.875	0.878	0.904	0.059	0.050	0.069
Item 9	156.870	43	0.000	0.878	0.882	0.907	0.058	0.049	0.068
Item 10	160.634	43	0.000	0.875	0.878	0.904	0.059	0.050	0.069

Table 6 presents the standardised and unstandardised regression weights of the gender-item paths in the MIMIC models. For example, model indicates that Item 1 maintains invariance of measurement across gender (standardised regression weight = 0.023; unstandardised estimate = 0.065; S.E. = 0.089; C.R. = 0.737; p = 0.461). C.R., the critical ratio, is the t value that is computed by dividing unstandardised regression weight by its standard error. C.R. values greater than |1.96| indicate statistical significance at $p < 0.05$ or a smaller p value. Overall, the MIMIC models confirmed the absence of DIF across gender for all items.

Table 6. Unstandardised and Standardised Regression Weights in the MIMIC Model

Item	Standardised estimate	Unstandardised estimate	S.E.	C.R.	p
1	0.023	0.065	0.089	0.737	0.461
2	0.024	0.070	0.091	0.771	0.441
3	0.034	0.064	0.061	1.045	0.296
4	-0.025	-0.043	0.059	-0.720	0.471
5	-0.023	-0.033	0.049	-0.663	0.507
6	-0.017	-0.039	0.072	-0.535	0.593
7	0.024	0.076	0.096	0.794	0.427
8	-0.009	-0.024	0.081	-0.295	0.768
9	-0.062	-0.059	0.030	-1.956	0.050
10	-0.008	-0.007	0.030	-0.225	0.822

Discussion

This study presents evidence of fairness in the PTE Academic Reading test, using two statistical procedures: the Rasch-based DIF analysis and the SEM-based MIMIC analysis. The Rasch-based DIF analysis found no UDIF across gender, but six cases of NUDIF interactions were identified in four items. Among the detected NUDIF pairs, three were in favor of low-ability females, disadvantaging high-ability males (Item 4 & 5) and high-ability females (Item 4). In addition, Item 1 exhibited NUDIF in favor of high-ability males while disadvantaging low-ability females; Item 1 also favored high-ability males compared with high-ability females. Items 4 and 9 have inverse NUDIF directions across high-ability females and low-ability females and, thus, cancel each other out and the detected NUDIF in Item 1 is negligible (DIF Contrast < 0.46). Item 4 also favors low-ability females while disadvantaging high-ability males, which cannot be canceled out. Overall, the significant and substantive NUDIF, which is not cancelled out, was observed in Items 4 and 5. The following post hoc content analysis discusses plausible reasons underlying the NUDIF.

Test takers are granted partial credits for each correctly ordered adjacent pair. Items 5 and 6 share the same test format: reordering sentences into a meaningful and logical paragraph. However, Item 6 does not induce any DIF across any subgroups, indicating that test format of reordering sentences might not be the cause of the observed DIF. Compared with Item 5, Item 6 is facilitated by more discourse markers and cohesive devices in the sentences, which can provide clues to the test takers to answer the Item correctly. In Item 5, the four options start with phrases as "These early faiths…, My study of the history…, Men and women…, This…". Reordering option B as the first sentence does not seem to be equally difficult, since the generic phrase "My study of …" is an indication of opening a new paragraph. It would not be too challenging to put either options A or D behind option C, as they start with demonstrative pronouns "These" and "This", indicating there must be some referents in the previous sentence. Option A, which starts with "these early faiths", could be directly connected with the "beliefs" in option C, while the demonstrative pronoun "This" in option D could be regarded as referring to what is stated in option C. Thus, both options A and D seem to be appropriate to be placed after option C, which could be confusing to some test takers. However, proficient readers would be more likely to find that D is a transitive sentence and serves to navigate the discussion towards statement A which is the key information the author wishes to convey. Successful completion of this item would require both grammatical and high-level topical knowledge to create a situation model or global representation as indicated in the integrated cognitive theory of comprehension and the construction integration model of comprehension (Kintsch, 1998). Failing to form a coherent situation model by relying on the grammatical competency and topical knowledge may result in guessing, which could help explain why the low-ability groups were advantaged by this item over the high-ability groups. (A NUDIF favors the low-ability female group over the high-ability male group.) This content analysis of Item 5 revealed that this pattern could be attributed to a guessing factor, since it is hard to make the correct choice between the correct answer and a distractor. One suggestion for cutting down or removing DIF could be adding stronger connective words between option C and D, so that the possibility of guessing factors could be reduced.

Item 4 is an MCQ question with two correct options out of five available options. A content analysis on Item 4 revealed that lexical complexity could be a facet contributing to the difficulty of this item. The item includes concepts such as "co-gastronomic movements", "academic food studies", "cultural anthropology", and "environmental anthropology" etc., which are less commonly used in non-technical reading passages. These phrases form strong distractors since they are not easy to be recognised and comprehended by the test takers, likely inducing the same type of guessing as in Item 5.

In some previous studies, female and male participants have been found to perform differentially in some high-stakes reading tests. In the PISA reading test, as an example, boys usually perform more poorly than girls due to potential cognitive differences (van Langen, Boskers, & Dekkers, 2006). T. I. Pae (2012) also discovered that items measuring the mood, emotion or tone could favor the female test takers, while the fill-in-the-blank items favor males. However, the findings of the present study seem not to support that there are evident differences in reading performance between females and males since only one out of ten items (10%) does not function equivalently across the genders.

As agreed by researchers, items flagged as DIF are not necessarily unfair or invalid (Camilli & Shepard, 1994; Wedman, 2017) and post-hoc studies would be needed to shed light on the main causes of DIF. Thus, as Harding (2011) has argued, DIF studies often result in further hypotheses and questions about the causes of DIF, and follow-up studies to examine the hypotheses and questions generated are necessary. It should be noted that the present study did not examine DIF across all test sections. Therefore, there is a possibility that the flagged DIF items could be cancelled out or exacerbated if the investigation of DIF was expanded to other test sections such as listening.

Overall, the SEM-based MIMIC analysis, on the other hand, demonstrated measurement invariance across female and male groups. In the MIMIC models, there was no evidence of gender-induced DIF, indicating that the reading ability of the test takers had been estimated reliably by the test items. One of the aims of this study was to determine whether the MIMIC analysis would produce a different result from the Rasch-based DIF analysis. As previously discussed, the Rasch-based DIF analysis detected both UDIF and NUDIF; the former was detected by comparing the probability of males and females of the same ability to answer each item correctly (Linacre, 2018b). To detect NUDIF, the Rasch-based DIF analysis divided groups (female and male) into smaller subgroups according to their ability (low-ability and high-ability subgroups).

NUDIF results can be visualised from the ICC generated by Rasch modeling. The displacement of difficulty estimates could be observed when the ICCs of two groups intercept with each other at two or more points. However, the visual representation has to be further supported by statistical analysis. On the other hand, MIMIC modeling is built on a different rationale that incorporates grouping variables instead of testing separate models for each group. Therefore, despite being a powerful tool to facilitate MI analysis for different background variables, it is not suitable for NUDIF analysis where groups (e.g., female or male) are divided into smaller subgroups by their ability (e.g., low-ability and high-ability subgroups). Collectively, the Rasch-based DIF analysis and SEM-based MIMIC model analysis reveal that the PTE Academic Reading test has achieved an acceptable degree of fairness as testified by both DIF and MI analyses and therefore, test takers' reading ability can be compared regardless of their gender.

Finally, there are several limitations in the study. The study is limited in scope since it only investigated the possible gender influence on test fairness. Other factors could be incorporated in the study, like academic background, computer literacy, and age, which have been found relevant in previous studies (Chubbuck, Curley & King, 2016; Pae, 2004; Geranpayeh & Kunnan, 2007). Moreover, future studies could extend to investigate multiple sets of test prompts in the same test to improve the accuracy of investigation.

Conclusion

The present study investigated test fairness in the PTE Academic Reading test by examining DIF and MI across gender with empirical evidence. Two statistical methods were adopted: the Rasch-based DIF analysis and the SEM-based MIMIC analysis. The results of these analyses were compared to establish measurement invariance, providing several implications for the PTE Academic Reading test and more broadly, language assessment.

First, test fairness evidence was found for the PTE Academic Reading test through the establishment of MI and lack of UDIF. Given its important role as a gate-keeper to higher-learning or immigration, the PTE Academic needs test fairness investigations to provide stakeholders with reliable evaluations and fair judgments. Compared with previous DIF studies on the PTE Academic (i.e. H. K. Pae, 2012; Song, 2014; Jin & Zhang, 2014), the present study is a focused investigation on only the reading test of the PTE Academic, offering a closer look at its fairness by combing two DIF methods. Moreover, the results of the study indicate there is a need to conduct similar studies on each subtest to provide a comprehensive quality check for the test.

Second, to our knowledge, it is the first study in language assessment that combines and compares the Rasch-based DIF analysis and SEM-based MIMIC analysis to conduct fairness investigations. The study has methodological implications for future research, which could adopt this combined research design rather than rely on only one technique (Teresi, 2006). Future research could also adopt other DIF detection methods such as Mantel-Haenszel and SIBTEST to compare their results, verify findings from analyses using different quantitative tools, and preclude the situation wherein DIF remains undetected due to the limitations of a particular quantitative method. The combined methods would also lend themselves to investigating DIF and MI in other high-stakes tests as part of their test validation endeavors.

In sum, the evidential basis for test fairness across gender is strongly supported by the UDIF and MIMIC results, despite the evidence yielded in the NUDIF analysis attenuating test fairness across "gender × ability" interactions. We provided some postulations concerning the causes of NUDIF, but it is crucial to perform further research to determine whether these postulations can be verified with larger pools of items. In the presence of the detected UDIF and NUDIF, the test developers might need to correct the test scores for DIF or search for evidence that could cancel out the effect of DIF.

Acknowledgements

This study was supported by a research grant from Pearson Education, UK, which was administered by the National Institute of Education of Nanyang Technological University, Singapore. The views and opinions expressed in this work are those of the authors and do not necessarily reflect the opinions of the above institutes.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91. doi.org/10.1111/j.1745-3984.1992.tb00368.x

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, *2*(2), 192-204. doi.org/10.1177/026553228500200207

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Arbuckle, J. L. (2006). *Amos 7.0 user's guide.* Chicago: Amos Development Corporation

Aryadoust, V. (2018). Using recursive partitioning Rasch trees to investigate differential item functioning in second language reading tests. *Studies in Educational Evaluation, 56*, 197-204. doi.org/10.1016/j.stueduc.2018.01.003

Aryadoust, V., Goh, C. M. C., & Lee O. K. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, *8*(4), 361-385. doi:10.1080/15434303.2011.628632

Banerjee, J., & Papageorgiou, S. (2016). What's in a topic? Exploring the interaction between test-taker age and item content in high-stakes testing. *International Journal of Listening*, *30*(1-2), 8-24. doi.org/10.1080/10904018.2015.1056876

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), S176-S181.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, *26*(4), 433-450.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press

Byrne, B. M. (2016). *Structural equation modeling with Amos: Basic concepts, applications, and programming* (3rd ed.). New York: Routledge.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park: SAGE Publications.

Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. *ETS Research Report Series*, 1992(2), i-143. doi.org/10.1002/j.2333-8504.1992.tb01495.x

Chubbuck, K., Curley, W. E., & King, T. C. (2016). Who's on First? Gender differences in performance on the SAT® test on critical reading items with sports and science content. *ETS Research Report Series*, *16*(2), 1-116. doi.org/10.1002/ets2.12109

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach. *ETS Research Report Series*, 1983(1), i-14. doi.org/10.1002/j.2330-8516.1983.tb00009.x

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355-68.

Engelhard, G. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* New York; London: Routledge

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113–148. doi.org/10.1080/15434300701375923

Gallagher, M. (2004). *A study of differential item functioning: Its use as a tool for urban educators to analyze reading performance* (Unpublished doctoral dissertation). Kent State University, Kent.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), 190-222. doi:10.1080/15434300701375758

Ginther, A., & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish language examination.

In Kunnan, J. (Ed.), *Validation in language assessment* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum

Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*(2), 163-180. doi.org/10.1177/0265532211421161

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

IBM Corp. (2016). *IBM SPSS Statistics for Windows* (Version 24.0) [computer software]. Armonk, NY: IBM Corp.

In'nami, Y., & Koizumi, R. (2011). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*(1), 131-152. doi:10.1177/0265532211413444

Jin, Y., & Zhang, X. (2014). *Effects of skill integration on language assessment: A comparative study of Pearson Test of English Academic and Internet-Based College English Test Band-6.* http://www.pearsonpte.com/research/Documents/RN_EffectsofSkillIntegrationonLanguageAssessmentOfPTEAcademic_2014.pdf.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631-639.

Lawrence, I. M., Curley, W. E. & McHale, F. J. (1988): *Differential item functioning for males and females on SAT verbal reading subscore items.* New York: College Entrance Examination Board. doi.org/10.1002/j.2330-8516.1988.tb00266.x

Linacre, J. M. (2018a). *Winsteps* [Computer program]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2018b). *A user's guide to WINSTEPS.* Chicago, IL: Winsteps.com.

Liu, J., & Dorans, N. J. (2016). Fairness in score interpretation. In N. J. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 77–96). New York: Routledge.

Lúcio, P. S., Salum, G., Swardfager, W., Mari, J. D. J., Pan, P. M., Bressan, R. A., ... & Cogo-Moreira, H. (2017). Testing measurement invariance across groups of children with and without attention-deficit/hyperactivity disorder: Applications for word recognition and spelling tasks. *Frontiers in Psychology*, *8*, doi: 10.3389/fpsyg.2017.01891

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), 89-114. doi:10.1177/026553220101800104

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*(3), 469-492.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge: Cambridge University Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York ; London: Guilford Press

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York ; London: Guilford Press

Langen, A. van, Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Educational Research and Evaluation, 12*(02), 155-177. doi.org/10.1080/13803610600587016

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute, 22*(4), 719-748. doi.org/10.1093/jnci/22.4.719

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. doi.org/10.1177/0265532211430367

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language Assessment.* Oxford: Oxford University Press.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Oxford: Blackwell Publishing.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*(2), 161-178. doi.org/10.1080/15434303.2011.565438

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational Measurement* (pp. 13-103). New York: ACE and Macmillan.

Messick, S. (1996). Validity and washback in language testing. Language Testing, 13(3), 241-256. doi.org/10.1177/026553229601300302

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36*(3), 217–232. doi.org/10.1111/j.1745-3984.1999.tb00555.x

Ockey, G. & Choi, I. (2015). Structual equation model reporting practices for language assessment. *Language Assessment Quarterly, 12*(3), 305-319, doi: 10.1080/15434303.2015.1050101

Oliveri, M. E., Lawless, R., Robin, F., & Bridgeman, B. (2018). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education, 31*(1), 1-16. doi.org/10.1080/08957347.2017.1391258

Pae, H. K. (2012). A psychometric measurement model for adult English language learners: Pearson Test of English Academic. *Educational Research and Evaluation, 18*(3), 211-229. doi.org/10.1080/13803611.2011.650921

Pae, T. I. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21*(1), 53-73. doi.org/10.1191/0265532204lt274oa

Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing, 29*(4), 533-554. doi.org/10.1177/0265532211434027

Park, G. P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly, 42*(1), 115-123. doi.org/10.1002/j.1545-7249.2008.tb00212.x

Pearson (n.d.). *PTE Academic – The English test that takes you places.* Last accessed Dec 23, 2018, from http://pearsonpte.com/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Raquel, M. R. (2019). The Rasch measurement approach to Differential Item Functioning (DIF) analysis in language assessment research. In V. Aryadoust & M. Raquel (Eds), *Quantitative data analysis methods for language assessment Volume I: Fundamental Techniques* (Chapter 5). New York, NY: Routledge.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36. doi.org/10.1177/0146621697211002

Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). *Papers in Language Testing and Assessment, 2*(2), 1-27.

Rouquette, A., Hardouin, J. B., Vanhaesebrouck, A., Sébille, V., & Coste, J. (2019). Differential Item Functioning (DIF) in composite health measurement scale: Recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. *PloS one, 14*(4), doi.org/10.1371/journal.pone.0215073

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12-29. doi.org/10.1177/026553229200900103

SAT Update. (2011, October 11). *SAT Critical Reading test development committee meeting* [PowerPoint slides]. New York, NY: College Board.

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.

Suh, Y., & Talley, A. E. (2015). An empirical comparison of DDF detection methods for understanding the causes of DIF in multiple-choice items. *Applied Measurement in Education, 28*(1), 48-67.

Song, X. (2014). *Test fairness in a large-scale high-stakes language test* (Unpublished doctoral dissertation). Queen's University, Canada.

Stricker, L. J., Rock, D. A. & Lee, Y. W. (2005). *Factor structure of the LanguEdgeTM Test across language groups* (TOEFL Monograph Series MS-32). Princeton, NJ: Educational Testing Service.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289-316.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. doi.org/10.1111/j.1745-3984.1990.tb00754.x

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*(3), 323–340. doi.org/10.1177/026553220001700303

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*(11). S152-S170.

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English academic: Building an assessment use argument. *Language Testing, 29*(4), 603-619. doi.org/10.1177/0265532212448619

Wedman, J. (2017), Reasons for gender-related differential item functioning in a college admissions test. *Scandinavian Journal of Educational Research*, doi: 10.1080/00313831.2017.1402365

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials.* Wilmington, DE: Wide Range.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147-170. doi: 10.1177/0265532209349465

Youn, S. J., & Im, S. (2016). Testing measurement invariance of an EAP listening placement test across undergraduate and graduate students. *Papers in Language Testing and Assessment*, *5*(2), 26-42.

Yoo, H., Manna, V. F., Monfils, L. F., & Oh, H. J. (2018). Measuring English language proficiency across subgroups: Using score equity assessment to evaluate test fairness. *Language Testing*. doi.org/10.1177/0265532218776040

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing*, *3*(1), 80–98. doi.org/10.1177/026553228600300104

Zeidner, M. (1987). A comparison of ethnic, sex and age bias in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing*, *4*(1), 55–71. doi.org/10.1177/026553228700400106

Zheng, Y., & De Jong, J. H. A. L. (2011). *Research note: Establishing construct and concurrent validity of Pearson Test of English Academic*. Retrieved from http://www.pearsonpte.com/research/Documents/RN_EstablishingConstructAndConcurrentValidityOfPTEAcademic_2011.pdf.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136-151. doi.org/10.1080/15434303.2014.972559

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1-28. doi.org/10.1111/j.1745-3984.1999.tb00543