Research Note: DIF investigations with Pearson Test of English Academic

Xiaomei Song, PhD Candidate
Queens' University, Canada

February 2014

1. Introduction

Fairness is of great importance in the milieu of high-stakes decision-making involving achievement, aptitude, admission, certification, and licensure. Fairness is threatened when tests yield scores or promote score interpretations that result in different meanings for different groups of test takers (Angoff, 1993). In the area of testing and measurement, attention of test fairness has been primarily given to avoiding bias in favour of, or against, test takers from certain groups in testing scores (i.e., ethnicity, gender, linguistic status, or socio-economic status) (Cole & Zieky, 2001). Bias research investigates construct-irrelevant components of a test that result in systematically higher, or lower, scores for identifiable groups of test takers (AERA, APA, & NCME, 1999).

The literature has used various approaches to examine test bias. Differential item functioning (DIF) is one of the most commonly used methods to detect potentially biased test items. It is a statistical procedure to judge whether test items are functioning in the same manner for different groups of test takers. Although DIF is a necessary but not a sufficient condition for test bias, it is a valuable tool to explore irrelevant factors that might interfere with testing scores, discriminate against certain groups, and produce inaccurate inferences. Its importance cannot be overemphasized, considering the potentially adverse influence that DIF items might exert on test taker groups.

Over the past decades, predominant research has been conducted examining psychometric properties of language tests (e.g., for reviews of DIF research in language testing, see Kunnan, 2000; Ferne & Rupp, 2007). These studies examined the effects of a variety of grouping variables such as gender (Aryadoust, Goh, & Kim, 2011; Breland & Lee, 2007; Takala & Kaftandjieva, 2000), language background (Elder, 1996; Kim & Jang, 2009), ethnicity (Freedle, 2006; Taylor & Lee, 2011), and academic background (Pae, 2004) on language test performance. In a special issue of *Language Assessment Quarterly* (2007), language testing researchers conducted DIF studies, including English language learners' test performance on math work problems (Ockey, 2007), English as a Second Language pragmalinguistics for test takers of Asian and European language backgrounds (Roever, 2007), and performance differences in a Cambridge ESOL test in terms of different age groups (Geranpayeh & Kunnan, 2007).

This study investigated DIF with Pearson Test of English Academic (PTE Academic). PTE Academic is a relatively new language test. It involves a growing number of test takers and other relevant stakeholders around the world (Pae, 2011). PTE Academic results are used for the purposes of admission, placement, and visa approval. This study examined gender effects on PTE Academic, with special interest on its integrated test formats. DIF detection has not been

conducted with integrated test formats. This study may have global implications in the area of language testing and assessment. PTE Academic context also presents a relatively novel perspective. By using the statistical DIF methods combined with content analyses of test items, the study will provide comprehensive and empirically-driven results regarding test validation and fairness.

2. Pearson Test of English Test and its integrated item formats

Pearson Test of English Academic is an international computer-based academic English language test, which aims to provide all test takers with equal opportunity to demonstrate their English proficiency (Pearson, 2008). PTE Academic measures test takers' language ability as required for entry to universities, higher education institutions, government departments, and other organizations requiring academic-level English. PTE Academic conducted its first field tests in 2007/2008. In 2009, Pearson formally launched PTE Academic, and its test centers were open in over 40 countries. PTE Academic scores are currently used by 3,000 programs across the world for admission purposes.

Based on the PTE Academic Offline Practice Test Overview (2011), there are three timed parts: Part 1 speaking and writing, Part 2 reading, and Part 3 listening. PTE Academic states that a fair test should be as relevant and authentic as possible, and one key feature to demonstrate the relevancy and authenticity is the use of integrated tasks (PTE Academic, 2012a). The use of integrated tasks reflects the real life language skills that students will need to apply in an academic environment. Among the three parts, for example, Part 3 includes 11 formats: summarizing spoken text, multiple-choice multiple answers, filling in the blanks, highlighting correct summary, multiple-choice single answer, selecting missing word, highlighting incorrect words, answering short questions, re-telling lecture, repeating sentence, and writing from dictation (PTE Academic, 2012a). While three formats - filling in the blanks, multiple-choice multiple answers, and multiple-choice single answer - are considered as independent format, the rest of eight are categorized as integrated format that requires the use of at least two skills. For instance, summarizing spoken text requires test takers to listen to a lecture, take notes, and then provide a written summary.

PTE Academic highlights the importance and potential advantages of using integrated tasks to increase authenticity and improve internal validity. However, limited research examines fundamental issues and evidence concerning integrated item formats (Weir, 2005).

3. Differential Item Functioning

Differential item functioning (DIF) is a statistical method to explore whether test items function differentially across different groups of test takers who are matched on ability. DIF exists when different groups of learners have differing response probabilities of either (a) successfully answering an item (i.e., in multiple choice) or (b) receiving the same item score (i.e., in performance assessment) (Ferne & Rupp, 2007; Zumbo, 2007). When group membership introduces a large DIF with consistent construct-irrelevant variance, it is generally

considered that the test measures something in addition to what it is intended to measure, and that the result is a combination of two or more than two measurements (McNamara & Roever, 1996). Alternatively, a large DIF would signal multidimensionality, suggesting that the test might measure additional constructs that function differently from one group to another (Gierl, 2005). According to Roussos and Stout (2004: 108), the general cause of DIF is that test items measure "at least one secondary dimension in addition to the primary dimension the item is intended to measure". Secondary dimensions are further categorized as either auxiliary dimensions that are part of the construct intended to be measured or nuisance dimensions that are not intended to be measured. Bias, thus, might occur if the existence of DIF is due to the situation where test items measure nuisance dimensions that are not relevant to the underlying ability of interest. What is concluded as potential bias or nuisance dimensions, may depend on subjective judgments, most often, the review of disciplinary experts.

The traditional, exploratory DIF approach was adopted in this study. Although it may be preferable to conduct DIF analyses based on substantive a priori hypotheses using the confirmatory approach, exploratory based DIF analyses are still common in the test development and evaluation process (Walker, 2011). Using an exploratory DIF analyses paradigm is often needed in practical DIF applications. According to Gierl (2005), the traditional, exploratory approach is conducted in two steps: statistical identification of items that favour particular groups, followed by a substantive review of potentially biased items to locate the sources of DIF. The traditional, exploratory approach has been used in the pervious empirical studies (Geranpayeh & Kunnan, 2007; Pae, 2004; Uiterwijk & Vallen, 2005).

To conduct the first step, several statistical procedures have been developed, including the Mantel-Haenszel method (MH), logistic regression (LR), the standardization procedure, and IRT (see a review by Clauser & Mazor, 1998). Developed by Shealy and Stout (1993), Simultaneous Item Bias test (SIBTEST) is one DIF procedure to explore how tests exhibit differential functioning at the item, as well as at the bundle[1], level toward different groups. Fundamentally, SIBTEST examines the ratio of the weighted difference in proportion correct (for reference and focal group member) to its standard error. DIF occurs 1) if an item is sensitive to both the primary dimension and a secondary dimension and 2) if the reference and focal groups that have been equated on the primary dimension differ in distribution on a secondary dimension (Russos & Stout, 1996). The SIBTEST procedure classifies items as having either "negligible (A-level, $|\beta| < .059$)" or "large (C-level, $|\beta| > .088$)", with "moderate (B-level)" being anything in between (Roussos & Stout, 1996).

SIBTEST has become one of the more popular DIF procedures for several reasons. First, SIBTEST has been proven to be a useful DIF procedure (Penfield & Lam, 2000; Walker, 2011). Zheng, Gierl, and Cui (2007) investigated the consistencies and effect size of three DIF procedures: MH, SIBTEST, and LR. Results showed consistent estimates on the magnitude and direction of DIF among the three DIF procedures. Second, SIBTEST uses a regression estimate of the true score based on iterative purification, instead of an observed score as the matching variable. As a result, test takers are matched on an estimated latent ability score, rather

[1] The term *bundle* refers to "any set of items chosen according to some organizing principle" (Douglas, Roussos, & Stout, 1996, p. 466). Gierl (2005) described four general organizing principles: content, psychological characteristics (e.g., problem-solving strategies), test specifications, and empirical outcomes.

than an observed score, which increases the accuracy of the matching variable. Third, SIBTEST can be used to explore differential functioning at the item and bundle levels. SIBTEST is one of a few procedures that can evaluate bundle DIF (DBF), and DBF provides increased power through more effectively controlled Type I error. Items with small but systematic DIF may very often go statistically unnoticed, but when combined at the bundle level, DIF may be detected (Roznowski & Reith, 1999; Takala & Kaftandjieva, 2000). Examining DBF becomes necessary to understand the influence of grouping variables on test performance, especially when important, although perhaps subtle, secondary dimensions associated with different bundles have been found in tests such as TOEFL (Douglas, Roussos, & Stout, 1996).

The substantive analysis is then conducted after the statistical DIF analysis. While DIF analyses identify differential performance across items, substantive analyses are required to determine the likely causes of the DIF and whether these causes are connected with the potential bias. The substantive analysis usually involves item reviews by subject area experts (e.g., curriculum specialists or item writers) in an attempt to interpret the factors that may contribute to differential performance between specific groups of test takers. A DIF item is potentially biased when reviewers identify the DIF sources that are due to components irrelevant to the construct measured by the test, placing one group of test takers at a disadvantage. Exploratory DIF analyses have been widely used in previous empirical studies, despite the situation that content analysis may not always provide conclusive answers regarding DIF sources, and reviewers cannot determine decisively that the existence of DIF and DBF is due to bias (Geranpayeh & Kunnan, 2007; Uiterwijk & Vallen, 2005).

4. Gender and test performance

Gender differences in cognition and learning have been long examined (Dennon, 1982; Hamilton, 2008). Numerous previous studies have been conducted to investigate gender differences in language proficiency performance, especially in terms of language skills, test content/topics familiarity, and test format/response types. Regarding language skills and ability, the findings differ significantly from conclusions "girls have greater verbal ability" (Cole, 1997, 11) to "there are no gender differences in verbal ability" (Hyde & Lynn, 1988, 62) to "women obtained lower means than men on the verbal scale' (Lynn & Dai, 1993, 462).

In terms of test content and topic familiarity, research generally found that males appear to be advantaged over females on physical, earth, and space science items in language tests (e.g., Brantmeier, 2003). Studies focusing on item format effect found that multiple-choice items seem to favour males and open-ended items tend to favour females (e.g., Bolger & Kellaghan, 1990). The possible reasons included the greater tendency of males in guessing the MC answers and higher quality in females' essay handwriting.

DIF methods provide an ideal way to examine gender effects on second language testing performance (Breland & Lee, 2007; Pae, 2004; Pomplun & Sundbye, 1999). Substantial DIF studies have been conducted to examine gender effects on language testing performance (e.g., Pae, 2004; Pomplun & Sundbye, 1999). By comparing groups that are matched on ability, DIF studies generally provide more detailed descriptions about the interactions between gender and language performance than the early non-DIF studies did. For example, Carlton and Harris (1992) examined gender DIF on the SAT using the MH procedure. Item level data from a total of 181,228 males and 198,668 females were analysed for DIF. The

results of the study showed that overall reading comprehension was differentially easier for the female group than the matched group of males, and males tended to perform better on antonyms and analogies than their female counterparts with equal ability. In the end, the authors concluded that DIF existed because women were likely to perform better on item types with more contextualized information. O'Neill, McPeek, and Wild (1993) extensively studied gender DIF across three forms of the Graduate Management Admission Test (GMAT). The study reported that reading comprehension items were differentially easier for males than for females matched on verbal ability, which seemed to be contrary to previous findings (e.g., Carlton and Harris, 1992). O'Neill's study also found that females tended to perform better on sentence correction items than males with equal verbal ability. Pae (2004) investigated the effect of gender on English language reading comprehension of the Korean National Entrance Exam for Colleges and Universities using the IRT Likelihood Ratio approach. The results of the study identified 28 gender DIF items out of a total of 38 test items at alpha level of 0.05 with half favouring males and the other half favouring females. Items classified as Mood/Impression/Tone tended to be easier for females, whereas items classified as Logical Inference were more likely to favour males regardless of item content. Further content analysis revealed that passage content was not a reliable factor that predicted interaction between gender and performance in reading comprehension, suggesting that future studies about gender effect on second language reading comprehension should consider item type as well as item content.

Empirical studies also discussed DIF cancellation. Takala and Kaftandjieva (2000) examined gender differences with a small sample of 475 examinees (182 males and 293 females) on the Vocabulary Test of the Finnish Foreign Language Certificate Examination. Using the IRT approach - the One Parameter Logistic Model - the results of the study showed that despite the fact that there were test items with indications of DIF in favour of either females or males, the test as a whole did not favour any gender groups. The number and magnitude of DIF items favouring females was almost equal to those favouring males, cancelling the effect of the DIF items. DIF cancellation has also been found and discussed in other studies (Pae, 2004; Roznowski & Reith, 1999).

Recently, attention has been given to DIF investigations on performance assessment. Breland and Lee (2007) examined gender differences on TOEFL CBT free-response writing examination performance. Forty-seven different prompts in Phase I and 87 prompts in Phase II were examined with a diverse population of test takers. A taxonomy of TOEFL writing prompts' characteristics was developed in Phase I using expert panel review and statistical analysis. It was found that the prompts having the largest gender differences tended to be about topics such as art and music, roommates, housing, friends, and children. The smallest gender differences tended to be associated with topics such as research, space travel, factories, and advertising. The Phase II analyses showed that the difference between the highest mean score and the lowest mean score was .30 standard deviations, which was considered to be relatively small. Nine prompts had mean score differences from other prompts exceeding .20 standard deviations. Almost all prompts analysed had statistically significant gender differences, but effect sizes were relatively small. Expert review of prompts at the extremes of difficulty and gender differences resulted in general agreement about what tends to characterize such prompts, but such characterizations did not always explain difficulty and gender differences. In the end, the researchers suggested a policy should be formulated for what levels of difference should result in prompts being dropped from active administration. In a similar vein, Ross and Okabe (2006) examined a 4-passage, 20-item reading comprehension test which was given to a stratified sample of 468 female and 357 male English as a foreign language (EFL)

learners. The test passages and items were also given to a panel of 97 in-service and preservice EFL teachers for subjective ratings of potential gender bias. The results of the actual item responses were then empirically checked for evidence of DIF using SIBTEST, the MH method, and the LR method. Concordance analyses of the subjective and objective methods suggested that subjective screening of bias overestimated the extent of actual item bias.

The above review of literature indicated that variation exists about the research design as well as the relationships between gender and language performance. This may be partially due to the fact that these studies investigated gender performance differences on tests that focused on various language skills and used different test format/responses types with different focuses. The current study intends to address two research questions:

1) How does the PTE Academic test exhibit differential functioning at the item and bundle level, if any, toward different gender groups?

2) What are the potential sources of the PTE Academic that content experts perceive to function differentially toward gender groups? Can these causes be linked to bias?

5. Method

5.1 Subjects and instrument

The study used item level data of PTE Academic for SIBTEST, which requires two groups of test takers to be tested with the same test items. According to Shealy and Stout (1993), SIBTEST can be used with sample sizes as small as 250 per group. After multiple oral and written communications with the PTE Academic team, only six items were identified to be close to this criterion, due to PTE Academic random composition of test forms from a large item bank. PTE Academic is a computer-based testing, and test items are rotated intact or partially over time. As such, it was unlikely to identify a large number of test takers being tested with the same test items at the time of the study. Within the current sample, there were 159 female and 241 male test takers. Test takers' background information and testing scores were collected. These test takers ranged from the youngest 17 to the oldest 55, with the average of 26.5 years old. The test takers came from different countries and they spoke various languages at home.

These six items were all dichotomously scored. None of them belonged to integrated item formats. The first two items were traditional, multiple-choice reading comprehension tasks selected from Part 2. The second two items asked test takers to select the best option after listening to the audio or watching the video and reading the alternatives from Part 3. The last two items required test takers to listen to the audio and complete the gapped written text by typing the missing word in each gap, also from Part 3.

5.2 Data analysis

Descriptive statistics was calculated to provide an overall picture of the data set. After that, the two-step exploratory approach was conducted: SIBTEST followed by content analysis carried out by content experts. SIBTEST was conducted only

at the item level due to the nature of the sample. This current study used female test takers as the focal group and male as the reference group. The second step of substantive analysis involved three content experts, all from the Pearson language testing team. Background information of the three content experts was unavailable to the researcher, so were the test items. The three content experts examined all the six items. In the answering sheet, they were asked to rate the suitability of the test items for each group using a questionnaire with a five-likert scale, and the format of the questionnaire followed the design of the DIF study by Geranpayeh and Kunnan (2007). Based on a scale from 1 (strongly advantage) to 2 (advantage) to 3 (neutral/neither advantage nor disadvantage) to 4 (disadvantage) to 5 (strongly disadvantage), the content experts rated each item and gave comments. In their answering sheet, the content experts were informed there were no right or wrong answers toward the results of content analysis. They were asked to consider various sources of potential bias including semantics, contents, vocabulary, pragmatics, context, historical background, or any other sources of potential bias.

6. Results

Table 1 reports the mean scores, standard deviation, skewness, and kurtosis for each group and overall. The descriptive statistics showed that male and female test takers performed similarly. Skewness and kurtosis values ranged between +1 and −1, indicating that the distribution of the data could be considered normal. Reliability estimates using Cronbach's alpha with the total scores were calculated. In general, these reliability estimates were high.

Table 1: Descriptive statistics

Grouping variable	N	Mean	SD	Kurtosis	Skewness
Female	159	49.03	13.17	.492	.392
Male	241	50.64	13.39	.308	.374
Total	400	50.00	13.31	.378	.341

Table 2 provides an overall description of the SIBTEST results at the item level. The gender SIBTEST analysis at the item level indicated that one item regarding multiple-choice listening comprehension showed B-level DIF favouring male (-.066). For security reasons, no further information about these test items was provided.

Table 2: SIBTEST results

Grouping variable	Item type	Beta Uni	Favouring
Gender	I1:Multiple-choices RC	-0.052	Female
	I2:Multiple-choices RC	-0.011	Female
	I3:Multiple-choices LC	0.017	Male
	I4:Multiple-choices LC	0.066	Male
	I5: Listen to the audio and type the missing word	-0.031	Female
	I6: Listen to the audio and type the missing word	-0.038	Female

The three reviewers examined whether these items showed potential bias towards females or males. They were advised to examine various potential sources including content, skill, format, semantics, vocabulary, genre, context, and test takers' experience. None of the three reviewers concluded that these items showed any potential bias toward gender groups. However, the three reviewers pointed out that some items might be biased toward different cultural groups. For example, the reviewers believed words and terms such as *Hugo Boss suits*, *federal and staffers*, and *the City Morning Herald* may bring difficulties for some cultural groups to accurately understand the meanings of the sentences or texts. One expert stated that, in a text related to Sydney, Sydney was not in the script so test takers had to figure out themselves that *the City Morning Herald* was a Sydney-based newspaper.

Table 3: Content analysis

Item type	R 1	R 2	R 3
I1:Multiple-choices reading comprehension	Neutral	Neutral	Neutral
I2:Multiple-choices reading comprehension	Neutral	Neutral	Neutral
I3:Multiple-choices listening comprehension	Neutral	Neutral	Natural
I4:Multiple-choices listening comprehension	Neutral	Neutral	Neutral
I5:listen to the audio and type the missing word	Neutral	Neutral	Neutral
I6:listen to the audio and type the missing word	Neutral	Neutral	Neutral

7. Discussion

This study investigated gender effects on PTE Academic. SIBTEST was used to explore the presence of DIF and quantified the size of DIF at the item level. When discussing these findings, it is important to keep in mind that differences do not reflect absolute group differences but rather relative performance discrepancies across items, after the groups have been matched for overall score. The current study identified one item, Item 4, favouring males at the B level. The result does not show systematic relationship between the DIF direction and item difficulty/item discrimination values. The low degree and magnitude of DIF was consistent with Pae's research (2012), which used IRT the Rasch model for DIF analyses with PTE.

While the SIBTEST results found one item with moderate DIF between gender and PTE Academic performance, the determination of test bias warrants further investigation through a content review of the test. The three reviewers from PTE Academic analysed all the items and found no potential bias towards gender groups. The reviewers were generally in concord that no evidence supported the notion that gender was a distinct construct to be measured in PTE Academic. However, they deemed whether and how contextualized manifestations of test scripts which may influence PTE academic performance needed further exploration. Although a Pearson internal study (Pearson, 2008) conducted sensitivity review with the full item bank of PTE Academic and changed or removed any instances of potential bias against, or in favour of, particular test

taker groups, what was regarded as potential cultural bias among the reviewers of this study seems to be different.

Cultural bias in language proficiency tests has long been discussed (Cheng & Henning, 1985; McGinley, 2002). For example, McGinley (2002) pointed out that some standardized tests used in the United States contain items that may be considered culturally bias. She mentioned *Woodcock-Johnson Revised*, a test which is replete with items describing nursery rhyme and American pop culture, was probably unfamiliar to learners from other cultures. A close examination of PTE Academic finds that PTE Academic highlights the importance of international academic English. PTE Academic adopts "an international flavor" of selecting texts and settings encountered in five countries: the United Kingdom, Canada, Australia, New Zealand, and the United States (Pearson, 2012). PTE Academic highlights the importance of international varieties of English and non-native accents since these varieties are highly relevant to today's modern international academic institutions. Test takers' awareness of multiple contextualized manifestations (e.g., content and accent) is encouraged and valued in PTE Academic. This intent is consistent with the current development of the use of international varieties as test input (Taylor, 2006). Whereas, such aim brings fairness concerns (Abeywickrama, 2013; Harding, 2012). Information is needed to provide as to how international academic English, which is defined as English used in the five countries, is equally represented in PTE Academic. Evidence is needed to show that tests and test items are produced to be fair and valid toward test takers who plan to study at different academic contexts across different counties and continents. To ensure equal treatment in the stage of item design and development, PTE Academic needs to consider whether multiple contextualized manifestations are all represented.

In the domain of language testing, these issues, as showed by this study, might bring difficulties in item design and development since reading/listening comprehension may consist of vocabulary, genres, and discourse that are context-situated. How to select and produce balanced test items and serve the multiple contexts and purposes in assessing test takers' ability presents challenges for PTE Academic developers. It is of importance to carefully explore these elements which may lead to problematic differential functioning and work to balance these elements as much as possible. Test developers need to consider a variety of variables in deciding what to test and how to test as well as in what quantity. Producing high quality tests, along with evidence that the test is of high quality, is a significant achievement in itself.

8. Conclusion

Since high-stakes tests play a significant role in decision-making, it is important to examine how tests function and what they really measure. This study found one item with moderate DIF toward gender groups. Content analysis by three reviewers suggested that the gender membership generally had minimal influence on the PTE Academic performance. As this study is exploratory by nature, more in-depth inspections using experimental methods are warranted. Special attention may be focused on the effects of contextualized manifestations in scripts on test performance. A balance is sought with regard to different groups in order to construct a language proficiency test that is fairly suited to all groups and individuals of test takers.

References

Abeywickrama, P. (2013). Why Not Non-native Varieties of English as Listening Comprehension Test Input? *RELC Journal*, 44 (1), 59-74.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC.

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, N.J.: Lawrence Erlbaum.

Aryadoust, V., Goh, C. C. M., & Kim, Lee. (2011). An investigation of Differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, 8, 361-85.

Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in Scholastic Acievement. *Journal of Educational Measurement*, 27, 165-174.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a Foreign Language*, 15, 1-27.

Breland, H., & Lee, Y.-W. (2007). Investigating uniform and non-uniform gender DIF in computer-based ESL writing assessment. *Applied Measurement in Education*, 20(4),377–403.

Carlton, S. T. & Harris, A. (1992). *Characteristic associated with Differential Item Functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparison (ETS-RR-64).* Princeton, NJ: Educational Testing Service.

Cheng, Z. & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests, *Language Testing*, 2, 155-163

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues & Practice*, 17, 31-44.

Cole, N. (1997). *The ETS gender study: How females and males perform in educational settings (ETS-RR-143).* Princeton, NJ: Edcuational Testing Service.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*, 38, 369-382.

Douglas, J., Roussos, L., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, 33, 465- 484.

Elder, C. (1996). The effect of language background on foreign language test performance: The case of Chinese, Italian, and modern Greek. *Language Learning*, 46, 233–282.

Ferne, T. & Rupp A. (2007). A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations. *Language Assessment Quarterly: An International Journal*, 4, 113-148.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, 73, 1-44.

Geranpayeh, A. & Kunnan, A. J. (2007). Differential item functioning in terms of age in the Certificate in Advanced English Examination. *Language assessment quarterly: An international journal*, 4, 190-222.

Gierl, M. (2005). Using a multidimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, 24, 3-14.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, 29, 163-180

Hyde, J. S. & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, 104, 53-69.

Kim & Jang, (2010). Differential Functioning of reading subskills on the OSSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. Language Learning, 59, 825-65.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), Fairness and and validation in language *validation in language assessment* (pp. 1-14). Cambridge, UK: Cambridge University Press.

Lynn, R. & Dai, X. (1993). Sex differences on the Chinese standardized sample of the WAIS-R. *The Journal of Genetic Psychology*, 154, 459-463.

McGinley, S. (2002). Standardized Testing and Cultural Bias. *ESOL Multicultural Newsletter*,

Kansas: Fort Hayes State University.
McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Oxford, UK: Blackwell Publishing.
O'Neill, K. A., McPeek, W. M., & Wild, C. L. (1993). *Differential Item Functioning on the Graduate Management Admission Test (ETS-RR-35).* Princeton, NJ: Educational Testing Service.
Ockey, G. J. (2007). Investigation of the validity of Math problems for English language learners with DIF. *Language Assessment Quarterly*, 4, 149-164.
Pae, H. (2011). *Differential item functioning and unidimensionality in the Pearson Test of English Academic*. PTE Report.
Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing*, 21, 53-73.
Pae, T. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, 32, 265-281.
Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and practices*, 19(3), 5-15.
Pomplun, M. & Sundbye, N. (1999). Gender differences in constructed response reading items. *Applied Measurement in Education*, 12, 95-109.
Pearson Test of English Academic (2008). PTE Academic sensitivity review project.
Pearson Test of English Academic (2011). PTE Academic Offline Practice Test Overview.
Pearson Test of English Academic (2012a). PTE Information.
Pearson Test of English Academic (2012b). Into the fourth year of PTE Academic—our story so far.
Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly*, 4, 165-189.
Ross. S. J. and Okabe, J. (2006) The subjective and objective interface of bias detection on language tests. *International Journal of Testing*, 6, 229-253.
Roussos, L, A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33, 215-230.
Roussos, L., & Stout, W. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook for social sciences* (pp. 107–115). Newbury Park, CA: Sage.
Roznowski, M. & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, 59, 248-269.
Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159-194.
Taylor, L. (2006). The changing landscape of English: implications for language assessment. *ELT Journal*, 60(1), 51–60.
Taylor, C. S. & Lee, Y. (2011). Ethnic DIF in reading tests with mixed item formats. *Educational Assessment*, 16, 35-68.
Takala, S. & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing*, 17, 323-340.
Uiterwijk H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, 22, 211-34.
Walker, C. M. (2011). What's the DIF? Why different item functioning analyse are important part of instrument development and validation. *Journal of Psychoeducational assessment,* 29, 364-76.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
Zheng, Y., Gierl, M. J., & Cui Y. (2007). 'Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST, and Logistic Regression procedures' Paper presented at the annual meeting of the NCME, Chicago, IL.
Zumbo, B. D. (2007). Three generation of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly: An International Journal*. 4, 223-233.