DIF investigations across groups of gender and academic background in a large-scale high-stakes language test

Xiamei Song
Georgia Southern University
Liying Cheng and Don Klinger
Queens' University

High-stakes pre-entry language testing is the predominate tool used to measure test takers' proficiency for admission purposes in higher education in China. Given the important role of these tests, there are heated discussions about how to ensure test fairness for different groups of test takers. This study examined the fairness of the Graduate School Entrance English Examination (GSEEE) that is used to decide whether over one million test takers can enter master's programs in China. Using SIBTEST and content analysis, the study investigated differential item functioning (DIF) and the presence of potential bias on the GSEEE with aspects to groups of gender and academic background. Results found that a large percentage of the GSEEE items did not provide reliable results to distinguish good and poor performers. A number of DIF and DBF functioned differentially and three test reviewers identified a myriad of factors such as motivation and learning styles that potentially contributed to group performance differences. However, consistent evidence was not found to suggest these flagged items/texts exhibited bias. While systematic bias may not have been detected, the results revealed poor test reliability and the study highlighted an urgent need to improve test quality and clarify the purpose of the test. DIF issues may be revisited once test quality has been improved.

Key words: Differential item functioning, test bias, language testing, content analysis, EAP/ESP

Xiamei Song, 2670 Southern Drive, Statesboro, GA 30460-8080. E-mail: sxmdaphne@yahoo.com

Introduction

High-stakes, pre-entry language testing is a predominate tool to measure test takers' knowledge and skills for the purpose of admission in higher education in China. These tests are used as a means to classify, select, and judge individuals. Given the important roles of high-stakes, pre-entry tests, there are concerted efforts to ensure that tests are fair to test takers. One example in obtaining empirical evidence of test fairness is to detect bias in the test in favour of or against test takers from certain groups (e.g., gender, linguistic, or ethnical status) that result in construct irrelevant differences in test scores (Cole & Zieky, 2001; McNamara & Roever, 2006). Differential item functioning (DIF) has become one of the most commonly used methods to judge whether test items function in the same manner for different groups of test takers. A similar procedure, differential bundle[2] functioning (DBF), provides a measure of performance differences across clusters of items, typically grouped by 'some organizing principles' (Douglas, Roussos, & Stout, 1996, p. 466). Although DIF and DBF are not sufficient to identify bias (Angoff, 1993; McNamara & Roever, 2006), they are valuable tools to explore irrelevant factors that might interfere with testing scores, discriminate against certain test taker groups, and produce inaccurate inferences.

Over the recent decades, DIF research has been conducted with various second language tests (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), language background (Kim & Jang, 2009), ethnicity (Taylor & Lee, 2011), age (Geranpayeh & Kunnan, 2007), and academic background (Pae, 2004) on language performance. As Ferne and Rupp pointed out (2007), DIF research in second language testing has 'only just begun to investigate DIF for language tests outside North America' (p. 144), and this is especially true for the Chinese context (Lei, 2007). As such, this study examined DIF and potential bias with one of the large-scale high-stakes language tests in Chinese higher education—the Graduate School Entrance English Examination (GSEEE) based on two grouping variables: gender and academic background. Fairness research is particularly important in the Chinese tertiary context in which there are an enormous number of test takers. Yet research using

[2] The term *bundle* refers to 'any set of items choose 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.

score-based empirical data to explore test fairness is not prevalent (Fan & Jin, 2012). DIF investigations may present a novel perspective to understand the fairness of the GSEEE in the context of Chinese higher education.

The GSEEE is designed and administered by the National Education Examinations Authority (NEEA) of the Ministry of Education of the People's Republic of China (Liu, 2010). The major purposes of the GSEEE are to measure English proficiency of test takers and to provide information for educational institutions to select candidates for their master's programs (He, 2010). According to its test specifications, the GSEEE examines test takers' linguistic knowledge in grammar and vocabulary, and skills in reading and writing (He, 2010). The total number of test takers for the GSEEE administration, for example, in 2011, reached approximately 1.51 million and the acceptance rate to enter master's programs was 32.75% (MOA, 2011). This test has significant consequences on millions of test takers who compete for graduate program admissions and educational opportunities.

Considering the demographic information of China and target test-taking population of the GSEEE, the study investigated DIF and DBF across groups of gender (female or male) and academic background (Humanities/Social Sciences or Sciences). First, a gender gap remains in the Chinese education and employment market (Postiglione, 2006). Females are less likely to receive higher education, especially in graduate schools (Liu & Li, 2010). Although gender differences on language performance have long been examined in literature (Cole, 1997; Kong, 2009), controversies exist regarding the interactions between gender and language performance, and it is unclear whether gender performance differences are due to test bias or ability differences. This study examined whether the GSEEE functioned differentially towards different gender groups and may bring advantages or disadvantages for one or other group, which, as a result, could lead to a gender gap in educational and employment opportunities. Second, the GSEEE is designed for all non-English major test takers in any areas of Humanities, Social Sciences, and Sciences (He, 2010). Literature has shown that test takers' academic background knowledge facilitates language performance (Kintsch, 1998; Tapiero, 2007). It is possible that the GSEEE differentially and unfairly favours test takers from certain academic background, which affects their opportunity to obtain master's education.

Examining how gender and academic background interfere with the GSEEE performance has important implications to explore whether test taker groups are

provided with equal opportunity to perform across a large country such as China. Since tests may not always be designed to keep the diversity of learner characteristics in mind, it is essential for test developers to monitor the test and examine its quality to see whether the test is fair to test taker groups (Geranpayeh & Kunnan, 2007). Specifically, the study addressed the following research questions:
(1) How do the GSEEE items and bundles exhibit differential functioning toward test taker groups of gender (female or male) and academic background (Humanities/Social Sciences or Sciences)?
(2) How do test reviewers perceive the possible causes of the differentially functioning GSEEE items and bundles? Can these causes be linked to potential bias toward test taker groups of gender and academic background?

Differential Item Functioning

Differential item functioning (DIF) is a statistical method to explore whether groups of test takers with equal ability have differing response probabilities of either successfully answering an item (e.g. , in multiple choice) or receiving the same item score (e.g. , in performance assessment) (Zumbo, 2007). The existence of DIF is due to the situation that test items measure 'at least one secondary dimension in addition to the primary dimension the item is intended to measure' (Roussos & Stout, 2004, p.108). Secondary dimensions are further classified as two types: auxiliary dimensions that are part of the construct intended to be measured and 'nuisance' dimensions that are not intended to be measured. If the existence of DIF is due to the situation that items measure nuisance dimensions, bias might occur. In this study, the traditional, exploratory DIF approach was adopted. 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 analysis paradigm may often be needed in practical evaluation of a test. The traditional, exploratory approach has been used in the previous studies, using various DIF techniques (Geranpayeh & Kunnan, 2007; Woo & Dragan, 2012).

The exploratory approach is often conducted in two steps: statistical identification of items that favour a particular group followed by a substantive review of potentially biased items to locate the sources of DIF (Gierl, 2005). To conduct the first step, a number of statistical procedures have been developed and tested such as the Mantel-Haenszel method (MH), logistic regression (LR), and IRT (see a review by Clauser & Mazor, 1998). This study used the technique of Simultaneous Item Bias Test (SIBTEST). Developed by Shealy and Stout (1993), SIBTEST is a nonparametric procedure to estimate DIF in an item or bundle of items. SIBTEST does not specify a formal model for item responding. It requires few assumptions and does not involve population parameters. Test takers are compared based on their membership in either the reference or focal group (e.g., male and female), where the suspicion is that the focal group might be disadvantaged on test items due to DIF. Items (bundles) on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest consists of those items suspected of measuring the primary and secondary dimensions; and the matching subtest contains items believed to measure only the primary dimension. SIBTEST has been proven to be a powerful DIF procedure (Penfield & Lam, 2000). It uses a regression estimate of the true score based on iterative purification instead of an observed score, which increases the accuracy of the matching variable. SIBTEST examines both uniform and non-uniform DIF. More importantly, SIBTEST is one of a few procedures that can evaluate bundle DIF, or DBF. Items with small but systematic DIF may very often go statistically unnoticed, but when combined at the bundle level, DIF may be detected due to the influence of local independence (Douglas, Roussos, & Stout, 1996; Roznowski & Reith, 1999; Takala & Kaftandjieva, 2000). In the GSEEE, all items are embedded in texts, and test takers answer item questions based on their understanding of those texts (bundles). As such, SIBTEST appears to be a useful tool to detect DBF given the feature of local independence of the GSEEE. The SIBTEST procedure identifies items as having either negligible (A-level, $|\beta| < .059$) DIF, moderate (B-level, $.060 < |\beta| < .087$) DIF, or large (C-level, $|\beta| > .088$).

After the statistical DIF analysis, the substantive analysis is conducted. The substantive analysis usually involves the review of items by content specialists or item writers in an attempt to interpret the factors that may contribute to differential performance between specific groups of test takers (Douglas, Roussos, & Stout, 1996). Substantive interpretations determine whether the item with DIF displays bias or impact. A DIF item is considered to be potentially biased when reviewers conclude that the DIF sources are due to irrelevant aspects, placing one group of test takers at a disadvantage. Considerable studies used substantive analysis to identify potential bias, despite the situation that reviewers may not always provide conclusive answers regarding DIF sources and they cannot determine decisively whether items with DIF display bias or impact (Geranpayeh & Kunnan, 2007; McNamara & Rover, 2006).

Gender and Test Performance

Gender differences in cognition and learning have long been examined (Dennon, 1982; Hamilton, 2008). Numerous early studies investigated gender differences in language proficiency performance, especially in terms of language skills and test format/response types. The findings vary considerably with respect to language skills and ability, from conclusions that 'girls have greater verbal ability' (Cole, 1997, p. 11) to 'there are no gender differences in verbal ability' (Hyde & Lynn, 1988, p. 62) to 'women obtained lower means than men on the verbal scale' (Lynn & Dai, 1993, p. 462). In terms of test content and topic familiarity, males appear to have an advantage on physical, earth, and space science items in language tests (Brantmeier, 2003). Studies focusing on item format effect generally concluded that multiple-choice (MC) items seem to favour males and open-ended items such as essay tend to favour females (Bolger & Kellaghan, 1990).

DIF methods provide an ideal way to examine gender effects on second language testing performance (Pomplun & Sundbye, 1999). Carlton and Harris (1992) examined gender DIF on the SAT. They found that overall reading comprehension was easier for the female group than the matched group of males, and males tended to perform better on antonyms and analogies than equally able females. O'Neill, McPeek, and Wild (1993) also extensively studied gender DIF across three testing forms of the GMAT. Their study reported that reading comprehension items were differentially easier for males than females matched on verbal ability, which seems to be contradictory to previous findings of Carlton and Harris (1992). Takala and Kaftandjieva (2000) examined gender differences with a small sample on a Vocabulary Test in Finland. Although there were test items that seemed to exhibit DIF in favour of either females or males, the test as a whole was not gender-biased since the observed differences in the test results remained even after excluding the DIF items. 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 (Roznowski & Reith, 1999). Extensive studies regarding gender effects on performance-based language tests have also been conducted, and they generally found females performed better than males on essays (DeMars, 2000). Breland et al. (2004) examined gender differences on TOEFL free-response writing examination performance. 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 were associated with topics such as research, space travel, factories, and advertising.

A few gender DIF studies have been conducted using large-scale test data in China. Lin and Wu (2003) examined gender differences on the English Proficiency Test (EPT). Using SIBTEST for DIF analyses and DIMTEST[3] for dimensionality investigation, they concluded that although the EPT did not demonstrate much gender DIF at the item level (2 items with C-level DIF and 11 with B-level DIF), DIF analysis with bundles of items provided evidence for a female advantage in listening, and a male advantage in cloze, and grammar and vocabulary. Kong (2009) analyzed gender differences with a small sample of test takers in the reading comprehension section of the Test for English Majors--Band 4 (TEM-4). Based on a SIBTEST analysis, two items with C-level DIF favouring females and one item with B-level favouring females were identified. Two passages showed C-level DIF at the bundle level, with one favouring females and one favouring males. Expert review of DIF and DBF concluded that the potential reasons for the DIF existence might be related with gender topics and problem-solving items. However, as these gender topics were 'not beyond the requirement of test specifications' (p. 17), Kong concluded that the existence of DIF as item impact and no test bias existed. Lastly, Lei (2007) examined the National Maculation English Test (NMET) and did not find any gender DIF across the 90 multiple-choice items and the essay item. Hence Lei concluded that the overall gender differences on the NMET were due to real differences in English language abilities between males and females.

The above review of literature indicates that variation exists about 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. DIF research is needed to investigate the gender effects on this specific large-scale high-stakes GSEEE, which has not been examined previously.

[3] DIMTEST examines the dimensional structure of a dataset and provides information about multidimensionality (Seraphine, 2000; Walker, Azen, & Schmitt, 2006).

Academic Background and Test Performance

The interactions between test takers' background knowledge and language proficiency, reading comprehension in particular, have been thoroughly studied in first language tests. Research papers consistently identify a facilitating effect of background knowledge on cognitive learning and reading comprehension theoretically and empirically (Kintsch, 1998; McNamara et al., 1996). The theory of Situation Models (Kintsch, 1998) describes how readers supplement the information provided by a text from their knowledge and experience in long-term memory to achieve a personal interpretation of the text that is related to this information in the text. First, in order to identify connections and effectively comprehend the text, the reader needs to process the text that consists of elements and relations within the text. Then, the reader must add nodes and establish links between nodes from his or her own knowledge and experience to make the structure coherent, complete gaps in comprehension, interpret information, and integrate all information with prior knowledge. Successful comprehension requires not only an adequate processing of the language (the semantics of words), but also the reader's familiarity with the situation described in the text that is gained through his or her interactions with the world and previous experiences. Needless to say, readers with abundant experience and domain knowledge tend to understand texts better than readers with little experience and domain knowledge (Kintsch, 1998; Tapiero, 2007).

In the second language area, limited studies have been conducted, and they consistently suggest there is a relationship between subject area knowledge and test performance (Chung & Berry, 2000; Krekeler, 2006). Hale (1988) examined whether students' academic discipline interacted with the text content in determining performance on the reading passages of TOEFL. Students in the Humanities/Social Sciences and Biological/Physical Sciences performed better on passages related to their own background knowledge than on other passages. The effect was significant for three of the four test forms; however, Hale concluded the effect was relatively small since the apparent subgroup advantage translated into 'only a few points in overall TOEFL scale score' (p. 59). Hale attributed this to the fact that TOEFL reading passages had been drawn from general readings rather than specialized textbooks. Examining the effect of background knowledge on the IELTS with test takers with a range of academic levels, Clapham (1996) also found that students generally performed significantly better on the reading modules in their own subject areas. However, the effect sizes were not sufficient to justify the ongoing provision of subject-area specific modules on the IELTS test.

Despite the identified links between academic background and language proficiency, few researchers have used DIF methods to explore such interactions. Pae (2004) used the MH procedure and the IRT likelihood ratio approach to investigate test performance between test takers in the Humanities and Sciences. The study found that seven items favoured the Humanities test takers, and 9 favoured the Science test takers. The preliminary content analysis of the test indicated that items dealing with science-related topics, data analysis, and number counting were differentially easier for the Sciences, whereas items about human relationships were differentially easier for the Humanities. It is unknown whether and how test items of the large-scale high-stakes GSEEE demonstrate DIF and potential bias toward different academic groups.

Method

This section describes participants, the version of the GSEEE administered in 2009, and data collection and analyses procedures. Before the study was conducted, ethics clearance had been received.

Participants

Applicants' background information and their GSEEE item-level data of the 2009 administration in one major university in South China were collected through one of the provincial NEEA branches. Among a random stratified sample of 13,745 applicants (test takers), 57.5% of the test takers were male and 42.5% were female. Approximately 8.4% of the test takers studied in the Humanities (e.g., literature, history, and philosophy), 16.3% in the Social Sciences (e.g., economics, psychology, and management), and 75.3% in the Natural and Applied Sciences (e.g., physics, chemistry, biology, and computer sciences). The information was similar to the demographic information of the overall GSEEE test-taking population (MOE, 2009). In addition, 25.4% of the test takers had graduated in previous years and 74.6% were in the last year of their undergraduate programs.

Three volunteer test reviewers, one male and two females, were also recruited for the study. The reviewers were purposely chosen based on their gender, age, work experience, and extensive knowledge of English teaching and testing. They were current university professors with extensive teaching experience in both undergraduate and graduate programs. They had all received professional training in Applied Linguistics with an emphasis on language testing and assessment. All of the test reviewers had participated in the test design of large-scale high-stakes

English tests in China, and two of them were involved in the GSEEE item writing. Before conducting the substantive analysis, the reviewers were briefed about the nature of the study, and they were given a copy of the testing paper and the items/texts needed for the content analysis. Prior to content analysis, all the test reviewers were given a letter of information that detailed their involvement in the study, and they signed a consent form.

The Graduate School Entrance English Examination

The 2009 administration of the GSEEE consisted of three sections (See Table 1). Section I, Cloze, consisted of the multiple-choice (MC) questions with 20 blanks in the text[4]. Section II, Reading comprehension (RC), included three parts: Parts A, B, and C. Part A contained 20 MC reading comprehension questions based on four reading passages on different topics, Part B was a text with five gaps where sentences were removed and test takers were required to match the most suitable option for each gap, and Part C was a text in which five sentences were required to be translated from English into Chinese. Section III, Writing, included two parts: Part A, a practical writing task, and Part B, an essay writing task. According to the test specifications, a practical writing task refers to a writing task used in everyday situations such as personal and business mails, memos, and outlines, whereas, an essay writing task requires test takers to produce a written argument or discussion on a given topic and give some examples to support their points. Test takers had to write about 100 words in the first task and 200 words in the second. Section I Cloze and Parts A and B in Section II, six texts in total, were dichotomously scored and weighted as 60 points out of a total of 100. The remaining three texts were polytomously scored.

[4] The GSEEE administered in 2009 consistently used the term *text* which consisted of more than one passage. The study used the same term throughout this paper.

Table 1. Description of the GSEEE Administered in 2009

Section	Part and item	Topic	Format	Score
I Cloze	Text (Items 1-20)	Animal intelligence	MC	10
II Reading	Part A Text 1 (Items 21-25)	Habits	MC	10
	Part A Text 2 (Items 26-30)	Genetic testing	MC	10
	Part A Text 3 (Items 31-35)	Education and economic growth	MC	10
	Part A Text 4 (Items 36-40)	The history of the New World	MC	10
	Part B Text (Items 41-45)	Theories of culture	Multiple matching	10
	Part C Text (Items 46-50)	The value of education	Translation	10
III Writing	Part A	White pollution	Practical Writing	10
	Part B	Internet: Connecting or separating all?	Essay Writing	20
Total				100

Data Analyses

To have an understanding of an overall picture of the GSEEE data set, descriptive statistics were calculated. Using Cronbach's alpha coefficients, reliabilities for the entire test and subtests were examined to provide an estimate of internal consistency. After that, the two-step exploratory approach was conducted (Gierl, 2005): SIBTEST that was used to identify DIF and DBF and the substantive analysis that explored the likely causes of DIF and DBF towards gender and academic background groups.

SIBTEST

SIBTEST was used for Section I Cloze and Parts A and B in Section II, 45 dichotomously-scored items in total. Poly-SIBTEST was used for 3 polytomously-scored items in Parts C in Section II and Section III. Female test takers and test takers from Humanities/Social Sciences were used as the focal group and male and Sciences as the reference group. The entire pool of 13,745 test takers was randomly reduced to 2000 for each reference and focal group. In order to guard against unrepresentativeness within the group of gender and academic background, an equal number of test takers with different characteristics were used to facilitate comparisons. Alternatively, when examining gender effects on the GSEEE, a stratified sample of 1000 female test takers from Humanities/Social Sciences and 1000 female test takers from Sciences were selected as the focal group; and a sample of 1000 male test takers from Humanities/Social Sciences and 1000 male test takers from Sciences were selected as the reference group. The sampling method was also applied to the investigation of academic background effects on the GSEEE. The stratified random sampling allows us to examine group effects with test takers from a diverse spectrum of characteristics and capture the major variations between the examined groups. Furthermore, when conducting SIBTEST, a standard one-item-at-a-time DIF analysis was performed in which each item was used as a suspect item and the rest serving as the matching criterion. Items displaying DIF were then removed from the matching criterion and DIF analysis was re-conducted. In terms of the DBF analysis, due to the nature that all dichotomously-scored items were embedded in the texts, DBF analysis was performed at the text level because each text apparently shared a common content theme. This bundling method is consistent with Gierl's recommendation (Gierl, 2005). DIF and DBF results were validated by multiple rounds of sampling with reference and focal groups.

Substantive Analysis

Substantive analysis was used to examine the reviewers' perceptions on DIF/DBF sources as well as whether these flagged items/texts were linked to the potential bias toward groups of gender and academic background. To complete this step, recorded telephone interviews were conducted with the three reviewers. Since individual reviewers with various backgrounds may interpret the sources of each DIF/DBF in different ways, a more comprehensive understanding of these flagged test items/texts might be achieved. The format of the substantive analysis was similar to that conducted by Geranpayeh and Kunnan (2007). The three participants were first asked to decide whether the flagged items/texts were likely to advantage/disadvantage test takers who were female or male and from Humanities and Social Sciences or Sciences background. They were asked to consider various sources of potential bias including context, semantics, content, vocabulary, pragmatics, or any other potential sources. They were then asked to rate the suitability of the flagged items/texts based on a scale from 1 (strongly disadvantage) to 2 (slightly disadvantage) to 3 (neither advantage nor disadvantage) to 4 (slightly advantage) to 5 (strongly advantage). Finally, the test reviewers were asked to explain their choices and make comments related to their choices.

Results

Descriptive Statistics and Test Evaluation

Table 2 reports the mean scores, standard deviation, skewness, and kurtosis for each group and overall. Results showed that female test takers outperformed males and test takers from Humanities/Social Sciences outperformed those from Sciences based on the total mean scores. Skewness and kurtosis values ranged between +1 and –1, indicating that the distribution of the data could be considered normal. Cronbach's alpha with each section and the total scores were calculated. In general, these reliability estimates were not high (α =0.53 for Section I; α =0.61 for Section II; α =0.65 for Section III; and α =0.71 for total), lower than the 0.70 standard (Pedhazur & Schmelkin, 1991).

Table 2. Results of Descriptive Statistics

Grouping variable		N	Mean	SD	Kurtosis	Skewness
Gender	Female	5678	49.52	10.55	.24	.39
	Male	7684	47.79	11.19	.09	-.41
Academic background	Humanities & Social Sciences	3300	49.00	11.57	.16	-.50
	Sciences	10062	48.40	10.75	.17	-.41
Total		13362	48.55	10.96	.17	-.41

Considering the low coefficient estimates, a follow-up investigation was conducted to examine item quality by using IRT-Bilog index. Generally speaking, the test showed a wide span of item difficulty with P-values (proportion correct) ranging from .09 to .85. Regardless, item discrimination values based on the point-biserial Pearson correlations were low, ranging from .02 to .35, with nearly two third below .20 (29 out of 45 MC items). In addition, two items—Item 12 and 43 had negative item discrimination values (-.07 and -.04 respectively), indicating good performers answered the items incorrectly or poor performers answered them correctly. These values show that the GSEEE test items did not function well to differentiate the high performers from low performers.

SIBTEST Results

Table 3 provides an overall description of the SIBTEST results at the item and bundle (text) level. The $|\beta|$ (Beta-uni) statistic was used as an effect size for gauging the magnitude of DIF (negligible DIF, moderate DIF, or large DIF). The DIF/DBF analysis was conducted with and without the two test items that showed negative discrimination values. Results found that the quantity and size of the flagged items and bundles remained even after excluding these two items (see Table 3).

Table 3. Results of SIBTEST analysis

Grouping variable	Section	Item/Bundle	Beta Uni with/without	P-value	Favouring
Gender	I Cloze	Item 14	-.075/-.075	< .01	Female
		Text (Items 1-20)	-.326/-.315	< .01	Female
	II RC	Part A Text 2 (Items 26-30)	.082/.081	< .01	Male
		Part A Item 37	.081/.081	< .01	Male
	III Writing	Essay writing	-.251/-.243	< .01	Female
Academic background	II RC	Part A Item 30	.060/.059	< .01	Sciences
		Part A Text 2 (Items 26-30)	.060/.060	< .01	Sciences

Regarding gender effects on the GSEEE, the SIBTEST analysis at the item level indicated that Item 14 in Section I showed moderate (B-level) DIF favouring females while Item 37 in Section II Part A showed moderate (B-level) DIF favouring males. Also, the essay writing task in Section III Part B showed large (C-level) DIF favouring female test takers. Specifically, Item 14 embedded in the text animal intelligence passage examines logical relationships; test takers were required to select the best word from four choices: (A) by chance (B) in contrast (C) as usual (D) for instance. Item 37 asks the test takers to determine the inferencing idea from one of the paragraphs in the text regarding the history of the New World. The essay writing task asks test takers to write an essay about 'Internet: Connecting or separating all?' based on a picture. With respect to the bundle DIF, results showed that the Cloze text regarding animal intelligence in Section I favoured females at a large level (large DIF) while Part A Text 2 regarding genetic testing in Section II favoured males at a large level (large DIF).

In terms of academic background effects on the GSEEE, the SIBTEST analysis at the item level found only one item, Item 30, which favoured the Sciences test takers. Item 30 asks test takers to identify an appropriate title for the text regarding genetic testing. When the bundle DIF came into play, Text 2 regarding genetic testing in Section II Part A showed moderate (B-level) DIF favouring the Sciences test takers.

Substantive Analysis Results

Three test reviewers examined what caused DIF/DBF with these flagged items/texts and whether these items/texts showed bias towards test taker groups of gender and academic background. Table 4 presents the reviewers' ratings of the flagged items/texts.

Table 4. Results of Content Analysis

Grouping variable	BIF Item/Bundle	Content analysis		
		Reviewer A	Reviewer B	Reviewer C
Gender	Item 14	3	3	3
	Section I Cloze (Items 1-20)	3	3	3
	Section II Part A Text 2 (Items 26-30)	3	4	3
	Part A Item 37	3	3	3
	Section III Essay writing	3	3	3
Academic background	Section II Part A Item 30	3	4	3
	Section II Part A Text 2 (Items 26-30)	4	4	3

Gender DIF and DBF

Item 14 was of average difficulty (62% correct) with good discrimination (0.40). The three test reviewers independently identified that Item 14 did not show bias towards male/female test takers. Item 14, which was embedded in the Cloze, examined inter- and intra-substantial logical relationships. The answer for this item was 'for instance' and this phrase was listed in the glossary of 5500 words that were required for the GSEEE test takers (NEEA, 2009). The three test reviewers' responses provided evidence of gender stereotyping. For example, they thought that the differential functioning of Item 14 between males and females was due to learner characteristic differences that females tended to pay much attention to details and often revisited the sentences/questions.

None of the three test reviewers rated the Cloze section (Items 1-20) about animal intelligence as exhibiting an advantage or disadvantage towards female or male test takers from its content, format, plot, or test takers' knowledge. Based on their teaching experience, the reviewers concluded that females generally outperformed males in identifying detailed information and noticing subtle changes in language. In comparison, males had the tendency to rush through the cloze items, often missing essential details that were necessary for a correct response.

Section II Part A Text 2 (Items 26-30) discussed the genetic testing of DNA. There was a gap in the conclusion of the three reviewers. Test Reviewer B felt that males performed better than their peers mainly because of males' advantages in science topic and content. Hence, this text may be slightly biased and favour males. In comparison, Test Reviewer A and C did not conclude any bias towards female or male test takers. They believed that discrepancies in English language proficiency were the major reason that caused the differential functioning of this text towards gender groups.

Item 37 was one of the more difficult items (42%) in the 2009 GSEEE with the low discrimination (0.21). None of the three test reviewers rated this item as biased towards female or male test takers. The item was embedded in Text 4 discussing the history of the New World. The primary target of this item was to correctly identify the main idea of one paragraph. The reviewers expressed a view that the existence of DIF was due to a general lack of knowledge and interest in history amongst females.

Regarding Section III the essay writing task that required test takers to write an essay regarding the Internet based on a drawing, none of the three test reviewers rated the DIF existence as biased towards female or male test takers. The test reviewers suggested that the differential functioning might be due to females' superior ability in the productive skills of speaking and writing. As such, the existence of DIF was considered as reflecting the groups' true differences on the construct. The reviewers also suggested that female test takers were more motivated and diligent than their male counterparts, and more likely to take the time to memorize words and sample essays.

Overall, it seems that the three reviewers focused on testing-taking techniques. Although during the process of content analysis they were asked to consider various sources of potential bias from various aspects such as context, semantics, content, vocabulary, and pragmatics, the reviewers attached much attention to test-taking strategies and preparation.

Academic background DIF and DBF

Item 30 was not hard (71% correct), but had a low discrimination index (0.11). The test reviewers had conflicting views about whether this item showed bias toward test takers with different academic backgrounds. Item 30 required test takers to identify an appropriate title for the text regarding genetic testing. Test Reviewer B concluded that Sciences test takers would benefit from their background knowledge slightly since they were generally more familiar with this topic. In contrast, the other two test reviewers, Test Reviewer A and C, felt no bias towards test takers from different academic backgrounds. They concluded that Item 30 was a global question since it asked test takers to generalize the title of the text, and it did not involve anything very complicated for test takers from the non-Sciences background.

Conflicting views among the reviewers also existed in terms of Section II Part A Text 2 (Item 26-30), which asked questions related to the DNA testing and its problems. Test Reviewer C did not think this text showed bias towards test takers from different academic backgrounds since the discipline-related topics such as the DNA testing fell within the scope of the GSEEE test specifications. In contrast, Test Reviewer A and B rated the text as favouring test takers from the Sciences background. They felt that the whole text was a passage of English for Specific Purpose (ESP) reading, which was beyond the scope of the GSEEE test specifications. The text was more than just general knowledge, and it was likely to favour those who had content knowledge. Therefore, they concluded that this text was biased slightly in favour of test takers from the Sciences background.

Discussion

This study investigated group membership effects on the GSEEE administered in 2009 regarding gender and academic background. Descriptive statistics found low reliability and discrimination values. Such results indicate the existence of flawed items in the GSEEE administrated in 2009. Low reliability estimates appear to fail significantly in terms of fairness because low reliability means that flawed decisions are being made about candidates. The GSEEE may underestimate (or overestimate) the ability of test takers who could perform better (or worse) and undermine its fairness claim for test takers. The GSEEE items that generated misleading and inaccurate information have the potential to lead to unfairness. Further, SIBTEST found the existence of DIF and DBF towards the groups of gender and academic background. Based on the results of content analysis, the input of the three test reviewers, though identifying a multitude of factors that they believed might have made these items/texts easier or harder for different groups, was not particularly informative. In the following, we will discuss the findings based on SIBTEST and content analysis.

SIBTEST Findings

SIBTEST was used to explore the presence of DIF and DBF and quantify the size of DIF and DBF. In terms of gender groups, the current study identified two items/texts favouring males at B level and three items/texts favouring females at C level. The results seem to show no systematic relationship between the DIF direction and item difficulty/item discrimination values. The results regarding the flagged items/bundles favouring males might be partially related with test content and topic familiarity. As the previous literature indicated, males tended to perform better in 'scientific-related content' than their matched female peers (Bond, 1993, p. 277). The study has also provided evidence that test format might be another reason to cause the differential functioning of the items/texts between male and female test takers. Compare with the practical writing task (10 out of 100 points), the effect of test format on gender performance differences is more evident in the essay writing since the essay writing took a larger percentage over the total score (20%). In this study, females performed better than males in essay writing, which is consistent with the conclusions from the previous studies (Bolger & Kellaghan, 1990; Pomplun & Sundbye, 1999). Moreover, female test takers performed significantly better than males on the Cloze. This result conflicts with Lin and Wu's study (2003) in which the bundles of Cloze favoured males slightly. As Cloze is a commonly used testing format in language testing (Bachman & Palmer, 2010), how gender interacts with the testing performance on Cloze certainly merits further research in the future.

Regarding the group of academic background, SIBTEST found that one item and one text functioned differentially towards the Sciences test takers. These results might be related to the content of the text, which focused on specific knowledge of Genetic Testing. This is consistent with results from the previous empirical studies which found students performed better on texts which related to their own background knowledge (Hale, 1988; Pae, 2004).

Substantive Analysis Findings

While the SIBTEST results found that interactions existed between gender, academic background, and the GSEEE test performance, the determination of test bias warrants further investigation through a content review of the test. Three experts in language testing examined the likely causes of the flagged items/texts and investigated whether these causes were linked to bias. Regarding gender DIF, the reviewers believed there were multiple factors associated with performance discrepancies. These factors include general, often gender stereotyped comments, e.g., that female test takers were advantaged in identifying information and subtle changes in language, in the productive skills of speaking and writing, and more highly motivated and diligent, while males were more knowledgeable in history and scientific-related topics. However, overall the three reviewers did not believe that gender played a major role in affecting test takers' admission status for master's programs in China.

Despite explicit explanation and their experience working as language experts in the universities and as high-stakes item writers, the reviewers did not seem to understand what bias meant. Additionally, the results of the content analysis reflect the challenge of using expert judges for bias analyses in DIF research, which is consistent with the previous studies (Geranpayeh & Kunnan, 2007; Uiterwijk & Vallen, 2005). The exploratory-based DIF analyses, though still common in test development and evaluation, generally fail to cogently explain causes beyond flagging the items that are potentially biased against certain groups.

Similar gender stereotyping occurred in relation to factors the reviewers associated with performance discrepancies towards groups of academic background. Further, the reviewers had different views regarding whether the flagged item and text were linked to potential bias. First, the reviewers seem to disagree whether the flagged item and text were disciplinary specific, or, whether the item/text examined general English or English for Academic Purposes (EAP)/English for Specific Purposes (ESP). To further understand the lexical component of the problematic text (Genetic Testing), a lexical text analysis was conducted using Vocabprofile (Cobb, 1999). Results found the passage did not seem very academic because of the low percentage of Academic Word List (AWL) words (5.5%). However, there were about 13% less frequent words in the Off-list category such as *paternity, prescription, kinship, genetic, saliva*, and *ancestor*. The top three key words for the proper comprehension of the text all fall into the scope of the Off-list category — *paternity, genetic*, and *ancestor*. Considering the Chinese national policies and institutional practices promoting the use of English as a teaching language in disciplinary courses (Du, 2003; Yang, 1996), test takers with Sciences background may have opportunities to learn these Off-list category words in their disciplinary courses. As such, test takers from Sciences background are more advantaged than those from Humanities/Social Sciences background in this text.

Second, the results suggest that the reviewers interpreted the GSEEE test specifications differently. There was no agreement among the three reviewers whether the GSEEE is designed to test General English, EAP, or ESP. While one reviewer believed the discipline-related topics such as the DNA testing fell within the scope of the GSEEE test specifications, the other two reviewers stated that ESP readings were beyond the scope of the GSEEE test specifications. According to the 2009 GSEEE test specifications (NEEA, 2009):

> The GSEEE is designed for non-English majors. Considering the practical purposes, test takers should have command of vocabularies related to one's profession or academic major as well as those involved in individual likes and dislike, life habits, and religion… Test takers should be able to comprehend not only books and newspapers in a variety of topics, but also literature, technical manuals, and introduction of products related to one's academic or professional area. (p. 2)

Hence, the purpose of the GSEEE is to examine test takers' knowledge and skills in General English as well as in EAP and ESP. However, this may create confusion and difficulties in test design and development because a great part of academic and professional vocabulary and reading comprehension consists of vocabulary, genres, and discourse that are discipline-specific and context-situated. The differences among General English, EAP, and ESP have long been discussed in language teaching, learning, and testing (Dudley-Evans & St John, 1998; Flowerdew & Peacock, 2001). While General English focuses on day-to-day communication, EAP and ESP are to meet specific needs of learners—academic learning and professional developments (Dudley-Evans, 1997). Thus EAP and ESP are centered on the language appropriate to these activities in terms of grammar, lexis, register, discourse, genre, and study skills. How to select and produce more balanced test items and serve the multiple purposes in assessing test takers' ability in General English as well as EAP and ESP presents a challenge for the GSEEE test developers.

Conclusion and Implications

As high-stakes, pre-entry tests play a significant role in decision-making, it is important to examine how tests function towards groups of test takers and what they really measure. This study found evidence of DIF and DBF on the 2009 GSEEE towards groups of gender and academic background. A review of the flagged items and tests by the three test reviewers identified a myriad of factors that potentially contributed to different performance of focal and reference group members who were matched on ability. Nevertheless, consistent evidence was not found to suggest these flagged items/texts exhibited bias. Such results indicate that the primary importance of improving the overall reliability of the test before further DIF studies should be attempted.

While systematic bias may not have been detected, the study provides important implications for the GSEEE test practices and English education in China as a whole. First, the study shows the urgency to improve item quality of the GSEEE. The GSEEE does not provide reliable results to estimate test takers' English proficiency through its test items. Given the low reliability and discrimination values, it is of paramount importance to ensure test quality so that individual test takers are provided with fair opportunities to perform. Although the NEEA claims to have established a quality control system and conducted test evaluation research (Liu, 2010), the public does not have access to its evaluation reports. How the test items with poor quality were addressed in score reports remains unknown. Because large-scale high-stakes language tests in China including the GSEEE have rarely been screened for item bias (Fan & Jin, 2012), the paper calls for moderation panels to conduct ongoing technical examinations. Second, there are some major implications for test developers and item writers. Test specifications should be clearly described in terms of the purposes of the test. The NEEA needs to revisit key elements in the GSEEE test specifications, such as test constructs, format, content, and how the choices on these issues may advantage or disadvantage certain groups or individuals of test takers. Proper training and item writing guidance should be provided to help item writers to be aware of potential bias issues in item design and content selection. The NEEA needs to perform DIF analyses and invite suitably trained reviewers to conduct fairness review to help ensure that test questions are fair to different groups. Evidence-based research (e.g., DIF techniques) will lead to improvements in test quality and establish a robust foundation and mechanism that improvements to test fairness can be built upon. Third, since significant group differences exist on the GSEEE items, there are implications for curriculum developers, university teachers, and students in teaching and learning in higher education. Curriculum developers may consider various topics in curriculum design in Chinese language education. Given the impact of tested language skills, topic familiarity, and test format on group performance differences, university instructors should adopt various teaching methods to encourage engagement of students with different backgrounds, and students should be encouraged to enhance their learning on certain aspects.

Since this is essentially an exploratory study, more in-depth, systematic inspections using confirmatory approaches are warranted in the future. The confirmatory approach for DIF analyses is theory-driven and allows for more thorough explanations of DIF (Ferne & Rupp, 2007; Sandilands et al., 2012). Rousos and Stout's (1996) two-stage multidimensionality-based confirmative DIF approach could provide direction for future studies. In addition, the use of multiple DIF procedures, such as Logistic regression, IRT, or MH alongside with SIBTEST, will be helpful to cross-validate statistical results and increase the certainty in identifying flagged items. As the groups of academic background (Humanities & Social Sciences or Sciences) were actually not equally represented in the real world population in China, inflation may exist and caution is needed to interpret the DIF results. In addition, the content analysis was post hoc; it would have been interesting to see which items were considered biased by content reviewers without guidance from the DIF analysis. The current study only focuses on one aspect of test fairness: potential bias in test design and development. Considering test fairness is 'an extraordinarily broad subject' (Willingham & Cole, 1997, p. 234), more research is needed to help to identify what aspects of the GSEEE may threaten fairness. Besides creating the best possible test items to eliminate bias, it is imperative to examine test fairness when the GSEEE is administered, scored, and used.

Acknowledgements

We would like to express our sincere gratitude to Prof. He and her team for their assistance in data entry. Thanks also go to the Ontario Graduate Scholarship program and Queen's University for funding the research. Finally, we want to thank Cathleen Benevento, Dr. Kathryn Hill, and two anonymous reviewers for their editing and revising work.

References

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.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice* (2nd ed.). Oxford:

Oxford University Press.

Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in

Scholastic Achievement. *Journal of Educational Measurement, 27*, 165-174.

Bond, L. (1993). Comment on the O'Neill and McPeek paper. In P. W. Holland & H.

Wainer (Eds.), *Differential item functioning* (pp. 277–281). Hillsdale, NJ: Lawrence Erlbaum.

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., Najarian, M., & Muraki, E. (2004). *An analysis of the TOEFL CBT*

writing prompt difficulty and comparability of different gender groups (ETS Research

Rep. No. 76). Princeton, NJ: Educational Testing Service.

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.

Chung, T. & Berry, V. (2000). The influence of subject knowledge and second language

proficiency on the reading comprehension of scientific and technical discourse. *Hong Kong Journal of Applied Linguistics 5,* 27–52.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge*

on reading comprehension. Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press. (Studies in Language Testing, Volume 4. Series Editor: Michael Milanovic.)

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify

differentially functioning test items. *Educational Measurement: Issues & Practice, 17*, 31-44.

Cobb, T. (1999). Breadth and depth of vocabulary acquisition with hands-on

concordancing. *Computer Assisted Language Learning 12*, 345-360.

Cole, N. (1997). *The ETS gender study: How females and males perform in educational settings*

(ETS-RR-143). Princeton, NJ: Educational Testing Service.

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

Measurement, 38, 369-382.

Dennon, D. (1982). Sex differences in cognition: A review and critique of the longitudinal evidence. *Adolescence, 17*, 779-788.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in*

Education, 13, 55-77.

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.

Du, W. (2003). On the Necessity of Teaching Specialized Course in English. *Journal of*

Shanxi Agriculture University (Social Science Education), 2(1), 72-74.

Dudley-Evans, T. and St John, M. (1998). Developments in ESP: A multi-disciplinary

approach. Cambridge: Cambridge University Press.

Fan, J. & Jin, Y. (2012). *Developing a code of practice for EFL testing in China: A data-based*

approach. Paper presented in LTRC, Princeton, NJ.

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.

Flowerdew, J., & Peacock, M. (Eds.). (2001). *Research perspectives on English for academic*

purposes. Cambridge, UK: Cambridge University Press.

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.

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.

Hale, G. A. (1988). Comprehension in the Test of English as a Foreign Language Student

major field and text content: interactive effects on reading. *Language Testing, 5*, 49-61.

Hamilton, C. (2008). *Cognition and sex differences*. New York: Palgrave MacMillan.

He, L. (2010). The Graduate School Entrance English Examination. In L. Cheng & A.

Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 145–157). New York: Routledge. *Psychological Bulletin, 104*, 53-69.

Krekeler, C. (2006). The effect of background knowledge revisited Language for special

academic purposes (LSAP) testing. *Language Testing, 23*, 99-130.

Kim, Y. & Jang, E. (2009). Differential Functioning of reading subskills on the OSSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning, 59,* 825-865.

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

University Press.

Kong, W. (2009). TEM-4 yuedu ceshi de DIF yanjou [DIF Study of Reading Module in

TEM-4], *Foreign Language in China*, 2009(1).

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

Lei, X. (2007). Shanghai gaokao yingyu fenshu de xingbie chayi he yuanying [Gender differences and their sources on the National Maculation English Test in the Shanghai area].*Shanghai Research on Education, 6*, 43-46.

Lin, J. & Wu, F. (2003). *Differential performance by gender in foreign language testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Liu, Q. (2010). The National Education Examinations Authority and its English language tests. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 29–43). New York: Routledge.

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

Liu, B. & Li, Y. (2010). Opportunities and barriers: Gendered reality in Chinese higher education. *Frontiers of Education in China. 5*, 197-221.

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

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.

Ministry of Education (2011). 2011 nian quanguo shuoshi yanjiushen zhaoshen kaoshi kaoshen jingru fushi [*National cut- scores of the 2011 administration*]. Retrieved on May20,2010 from http://yz.chsi.com.cn/kyzx/kydt/201203/20120330/296907941.html

National Education Examinations Authority (2009). *GSEEE Syllabus.* Beijing, China: Higher Education Press.

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.

Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing,*

21, 53-73.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates

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.

Postiglione, G. A. (2006). School and inequality in China. In G. A. Postiglione (Ed.),

Educational and social change in China: Inequality in a market economy (pp.3-24). NY:

M.E. Sharpe, Inc.

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.

Tapiero, I. (2007). *Situation models and levels of coherence: Toward a definition of*

comprehension. New York: Lawrence Erlbaum.

Takala, S. & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary

test. *Languaeg Testing, 17*, 323-340.

Taylor, C. S. & Lee, Y. (2011). Ethnic DIF in reading tests with mixed item formats.

Educational Assessment, 16, 35-68.

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

Willingham, W. & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, New Jersey:

Lawrence Erlbaum.

Woo, A. & Dragan, M. (2012). Ensuring validity of NCLEX with different item functioning analysis. *Journal of Nursing Regulation, 2,* 29-31.

Yang, X. (1996). Exploring the teaching of disciplinary courses in English. *Jinlin*

Education & Science, 2, 44-45.

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.