Journal of Language Testing & Assessment (2020) Vol. 3: 5-19 **DOI: 10.23977/langta.2020.030102**
Clausius Scientific Press, Canada **ISSN 2616-2938**

Gender-related Differential Item Functioning Analysis on an ESL Test

Don Yao[1, a, *], Kayla Chen[2, b]

[1]*Language Assessment Seminar Research (LASeR) Group, Department of English, Faculty of Arts and Humanities, University of Macau, Macau, China*
[2]*Xingang Middle School, Huangpu District, Guangzhou, China*
[a]*yb87710@um.edu.mo*, [b]*15521319421@163.com*
**Corresponding author*

Keywords: Differential Item Functioning, Gender, Mantel-Haenszel Method, Young Language Learners, Bias

Abstract: Differential item functioning (DIF) is a technique used to examine whether items function differently across different groups. The DIF analysis helps detect bias in an assessment to ensure the fairness of the assessment. However, most of the previous research has focused on high-stakes assessments. There is a dearth in research that laying emphasis on low-stakes assessments, which is also significant for the test development and validation process. Additionally, gender difference in test performance is always a particular concern for researchers to evaluate whether a test is fair or not. This present study investigated whether test items of the General English Proficiency Test for Kids (GEPT-Kids) are free of bias in terms of gender differences. A mixed-method sequential explanatory research design was adopted with two phases. In phase I, test performance data of 492 participants from five Chinese speaking cities were analyzed by the Mantel-Haenszel (MH) method to detect gender DIF. In phase II, items that manifested DIF were subject to content analysis through three experienced reviewers to identify possible sources of DIF. The results showed that three items were detected with moderate gender DIF through statistical methods and three items were identified as possible biased items by expert judgment. The results provide preliminary contributions to DIF analysis for low-stakes assessment in the field of language assessment. Besides, young language learners, especially in the Chinese context, have been drawn renewed attention. Thus, the results may also add to the body of literature that can shed some light on the test development for young language learners.

1. Introduction

Test fairness is always closely related to test validity and validation in the field of language assessment (Kunnan, 2010). In accordance with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), test fairness refers to “examinees of equal standing with respect to the construct of the test is intended to measure should on average earn the same test score, irrespective of group membership” (p. 74). In the process of test validation, one way to determine

the existence of bias in a test is to examine the group differences on test performance. Differential Item Functioning (DIF) is a technique that identifies these items that function differently favoring a subgroup of test takers. DIF studies help detect bias in a test and ensure a test is fair to test takers. As Walker (2011) suggested, DIF analysis is a significant part of test development and validation. In other words, a large number of items exhibiting DIF would threaten the construct validity and fairness of a test.

For high-stakes assessments, test takers' performances may bring about great impact on their life at the macro level, such as university admission. However, the essentiality of students' performance on low-stakes assessment for the improvement of curriculum and instruction at the micro level also attaches great importance (Martinková et al., 2017). In addition, gender difference in test performance is always a particular concern for researchers to evaluate whether a test is fair or not in terms of gender (Lin & Wu, 2004). There is an increasing attention has been paid to gender-related DIF studies on language tests (e.g., the English Proficiency Test [EPT], the Graduate School Entrance English Examination [GSEEE], the General Scholastic Ability Test [GSAT], and General English Proficiency Test [GEPT]) in Chinese speaking areas especially over the last decade.

To bridge the gap and expand the body of research on gender-related DIF, the present study investigated whether the test items of the General English Proficiency Test for Kids (GEPT-Kids) are free of bias in terms of gender differences. The GEPT-Kids was developed in 2015 by Language Training and Testing Center (LTTC). It is not a high-stakes but large-scale assessment for elementary school students in Taiwan, which plays a significant role in providing diagnostic feedback and reference for future teaching and learning. A mixed-method sequential explanatory research design was adopted with two phases. In phase I, the Mantel-Haenszel analysis was used to detect DIF and potential bias in the GEPT-Kids. In phase II, content analysis (i.e., expert judgment) was employed to identify possible sources of DIF and further elaborate on the reasons that may cause the DIF. The study, therefore, attempted to identify the existence of substantial DIF favoring males or females and examine potential sources of biased items through content analysis.

The current study may carry profound implications to the field of language assessment. To begin with, gender DIF analysis was conducted on a low-stakes assessment (i.e., GEPT-Kids) which contributes to fairness of the test. It plays an important role in English learners' future English learning. Any test item with bias would affect the real performance of test takers, which would negatively influence their recognition of their levels and teaching decisions made by teachers or parents. Also, research on gender DIF in GEPT-Kids provides reference for test developers to select test materials and design the test display. Besides, gender-related DIF studies are not common in language tests for young learners and particular in the Chinese speaking context. Only one study conducted DIF analysis on a primary English test of a province in China (Zhu et al., 2017). Hence, the present study might add to the body of literature that can make a contribution to test development for young English learners (e.g., McKay, 2006; Bailey, 2008; Hasselgreen & Caudwell, 2016; Papp & Rixon, 2018).

2. Literature Review

2.1. Previous DIF Research

Test fairness is imperative to both test makers and users, which is closely related to test validity and validation (Kunnan, 2007). However, it might be influenced by bias in test items or test levels (Jiao & Chen, 2014). Especially, the ability of a group can be appraised inaccurately by test bias through group comparisons. As mentioned earlier, DIF is a technique that identifies items function differently favoring a subgroup of test takers. It exists when test takers from different groups who have equal ability but do not have equal probability of successfully answering an item or receive the

same item score (Shepard et al., 1981; Zumbo, 2007). There are two types of DIF, i.e., uniform DIF and non-uniform DIF, need to be considered during DIF analysis. Uniform DIF affects students at all levels of the total score while non-uniform DIF affects students in specific ranges of the total score inconsistently (Martinková et al., 2017). The fundamental principle of DIF that test takers who master the same amount of knowledge of a topic are supposed to have equally good performance on an item testing that topic despite gender, race, or ethnicity (O'Neill and McPeek, 1993). If a test is free of bias, it is expected to shun significant score differences among test takers. According to Shepard et al. (1981), multidimensionality leads to the unequal probability of answering items correctly. Generally, the primary dimension of a test is to assess one latent construct. Nevertheless, items displaying DIF might assess at least one additional dimension. In other words, an item might show DIF against a group if test takers have less ability on the additional dimension.

Researchers have believed that DIF detection methods are feasible to reveal factors that account for group differences (Ahmadi & Bazvand, 2016; Pae, 2004). It is also worthy to note that items flagged as DIF do not necessarily indicate bias or unfairness. It only represents that those items should be carefully examined in order to avoid construct irrelevance or construct underrepresentation (Messick, 1989; Banerjee & Papageorgiou, 2016). Also, factors contributing to differential group performances may vary, e.g., test content, or test format (Kunnan, 2017). DIF analysis has been conducted in numerous language testing programs across various subgroups of test takers (e.g., gender, ethnicity, academic majors, native languages, age) after matching on language ability. A full list of DIF-related studies has been concluded by Kunnan (2017, see Table 1). It was found the interest in gender and DIF could be dated back to 1986. And until now, it is still a prevalent topic in the arena of language assessment.

Table 1. Empirical Studies in Language Testing Focusing on Test Bias (from 1980).

Authors and Year of Study	Specific Focus
Swinton & Powers (1980)	L1 language
Alderman & Holland (1981)	L1 language
Shohamy (1984)	Test method
Alderson & Urquhart (1985)	Academic major
Chen & Henning (1985)	L1 language
Zeidner (1986, 1987)	Gender and minorities
Hale (1988)	Major field and test content
Oltman, Stricker & Barrows (1988)	L1 language
Kunnan (1990, 1992, 1995)	L1 language and gender
Sasaki (1991)	L1 language
Shohamy & Inbar (1991)	Question type and listening
Ryan & Bachman (1992)	Gender
Brown (1993)	Tape-mediated test
Ginther & Stevens (1998)	L1 language and ethnicity
Norton & Stein (1998)	Text content
Brown (1999)	L1 language
Takala & Kaftandjieva (2000)	L1 language
Lowenberg (2000)	Different Englishes
Kim (2001)	L1 language
Pae (2004a, 2004b)	Academic major and gender
Pae & Park (2006)	Gender
Abbott (2007)	L1 language
Ockey (2007)	L1 language
Roever (2007)	L1 language
Geranpayeh & Kunnan (2007)	Age
Allalouf & Abramzon (2008)	L1 language
Kim & Jang (2009)	L1 language
Aryadoust, Goh & Kim (2011)	Gender
Aryadoust (2012)	Age
Harding (2012)	L1 language
Banerjee & Papageorgiou (2016)	Age
Grover & Ercikan (2017)	Gender and socioeconomic status

Note. Adapted from *Evaluating Language Assessment* (Kunnan, 2017, p. 173)

2.2. Gender-related DIF Research of High-stakes Assessments

A sizable number of early studies on gender and language test performance have been conducted to examine whether males and females have different performance in various test tasks. Whether such difference is determined by true gender ability or test item bias has impact on test fairness should be analyzed (Fernandes, 2015; Kunnan, 2007). Overall, in terms of item content, related research reports that males tend to outperform females at the aspect of items concerning science while females have advantage over males in regard to items about social sciences, aesthetics, philosophy, human relationships, and arts or humanities (Brantmeier, 2003; Carlton & Harris, 1992; Kunnan, 1990; O'Neill & McPeek, 1993; Scheuneman & Gerritz, 1990). Additionally, researchers have starched the surface on the relationship between item types and gender DIF as well. Reportedly, males perform better on multiple-choice items and females are favored by open-ended items, e.g., essay writing (Bolger & Kellaghan, 1990; DeMars, 2000).

Gender DIF is one of the most popular topics which has been investigated by diverse DIF detection methods. Accordingly, a number of second or foreign language testing programs have been examined such as the Test of English as a Foreign Language (TOFEL) (Ryan and Bachman, 1992), the Graduate Record Exam (GRE) (Scheuneman & Gerritz, 1990) and the Graduate Management Admission Test (G-MAT) (O'Neill et al., 1993). Gender DIF researchers also made an effort on tests for subskills, e.g., listening and reading (Aryadoust et al., 2011; Pae, 2004, 2012; Park, 2008). With regard to Liu and Li (2010), males are more likely to pursue higher education in China. And gender difference on language tests within Chinese-speaking context has gradually been paid more attention.

Kunnan (1990) conducted a study to identify DIF among gender groups in the fall 1987 version of English as a Second Language Placement Exam (ESLPE) through the one-parameter Rasch model. A total of 23 items exhibited gender DIF from all sections of the test. Twenty items favored males while three favored females. It was reported that potential sources for listening and reading items with DIF favoring males were content of test passages and male test takers' academic major background. Males had an advantage on test content related to business, archeology, and aerospace engineering. In terms of grammar, vocabulary and writing error detection items were in favor males and the three DIF items were in favor of females. However, their potential source for DIF could not be hypothesized.

Ryan and Bachman (1992) detected gender bias in TOEFL and First Certificate of English (FCE) with content categorized as structure, vocabulary, and reading by Mantel-Haenszel (MH) procedure. It was concluded that four items favored males while two were against females in TOEFL and two items favored males and females respectively in FCE, which all demonstrated moderated DIF. The results suggested that there was no distinct difference between males and females' performance. Authors also called for closer examination of the relationships between DIF and item content.

Lin and Wu (2003) explored the existence of gender bias in the English Proficiency Test (EPT). Among the 120 multiple-choice items analyzed by SIBTEST, thirteen test items exhibited DIF including five favoring females and eight favoring males. They arrived the conclusion that there was not much gender DIF shown at the item level. The further analysis by DIMTEST showed that females were favored by the bundle of listening comprehension test items while males were favored by the bundles of grammar, vocabulary and cloze, which is consistent with overseas gender DIF studies (Pae, 2004, 2012).

Wu (2009) investigated DIF in gender in the 2006 General English Proficiency Test (GEPT) listening tests by MH method. The result showed that only 3% of the test items displayed medium DIF. There was no evidence demonstrating that test items favored females in GEPT listening, even

though female test takers acquired higher scores than males in the section. Further research was also called for enhancing test fairness within the Taiwanese context.

Song et al. (2015) evaluated the fairness of the Graduate School Entrance English Examination (GSEEE) in China with the use of DIF detection methods. They took gender as one of the grouping variables by means of SIBTEST and content analysis. DIF and differential bundle functioning (DBF) were spotted towards gender groups. It was reported that two DIF flagged test items were easier for males, whereas three items favored females. Females overperformed than males in cloze, which is conflicted with Lin and Wu's (2003) research results. Motivation and learning styles were identified as factors that resulted in discrepant performance of gender groups.

Lorenz (2016) conducted a gender DIF research employing item response theory (IRT) for the A-level examinations in English as foreign language. Data were collected from 1,136 student participants. It turned out that only three test items were detected DIF. One favored females and two were easier for males. The item in favor of females required students to analyze grammar of the text. Lorenz (2016) pointed out that female test takers might have higher abilities in analyzing the text and, thus, had higher probability to gain higher scores. The topic of the two items favoring males were both related to terrorism, police, and technical surveillance, which was considered as more attractive and interesting for male test takers.

All the aforementioned research has provided a solid foundation both theoretically and practically in conducting DIF research. It is not difficult to find that the previous research has only spotted on high-stakes assessments. However, the fairness issues in low-stakes assessments also attach great importance in the field of language assessment (Wise & DeMars, 2005, Martinková et al., 2017). The next sub-section, therefore, reviews gender-related DIF research in low-stakes assessments.

2.3. Gender-related DIF Research of Low-stakes Assessments

To date, DIF analysis for evaluating fairness of tests and reasons behind the differential test performance between subgroups mostly exists in the high-stakes testing world but rare DIF analysis has been conducted on low-stakes tests (Martinková et al., 2017). High-stakes assessments are always emphasized because they bring about more significant consequences compared with low-stakes assessments, e.g., university admission or company recruitment.

But Martinková et al. (2017) claimed that DIF research in low-stakes assessment also merits attention. They, therefore, tested for gender bias in the Homeostasis Concept Inventory (HCI) and a simulated dataset inspired by the Graduate Management Admission Test (GMAT). The result of the first test by MH, linear regression (LR) and IRT showed the lack of test items with serious DIF even though the achievement gap between males and females was significant. In the second study, analysis on the designed data by same methods used in the first test reported two items with DIF even when the two groups had exactly the equal distribution of total scores. The authors argued that DIF analysis was much more useful to detect test bias in lieu of the observation of total scores when comparing group performances. It was necessary when developing and designing low-stakes assessment.

Given the scarcity in research on DIF and low-stakes assessments and the prevalence of the factor gender in the field of language assessment, two research questions were articulated:

- RQ1: Did GEPT-Kids test items display DIF in favor of male or female test takers after matching on ability?
- RQ2: Did content analysis of DIF items indicate possible bias in GEPT-Kids in favor of male or female test takers?

3. Methodology

3.1. Participants

There are two groups of participants in the present study. First, a total of 492 (M=226, F=267) elementary students from Beijing, Shanghai, Guangzhou, Macau, and Hong Kong took part in the GEPT-Kids test. They are all Grade 5 or Grade 6 students. Their first language is Chinese Mandarin or Cantonese and their second language is English. They have been learning English since Grade 1. Second, three Ph.D. students (i.e., content reviewers) were recruited for both rating and content analysis (or expert judgment). They all major in English linguistics and are now studying at a public university located in Macau. They all have been learning English for more than 15 years and have experience in teaching English before. They helped rate the test paper first, followed by reviewing the test paper of GEPT-Kids and elaborating on the possible reasons that may cause DIF.

3.2. Instruments

3.2.1. GEPT-Kids Test Paper

The GEPT-Kids test paper was developed by the LTTC, Taipei. Colorful pictures are illustrated in the test along with test items. The test booklet in the current study consists of two sections, i.e., Listening and Reading (see Table 2). Both of the listening and reading sections are delivered in paper-and-pencil mode. There are 55 test items in the test booklet and all the questions are in multiple-choice (MC) formats and scored dichotomously. The full score of the test weighs 55 points.

The listening section has four parts with 25 test items. The first two tasks are yes or no questions. As for the first task, five pictures are displayed and test takers are required to indicate whether each short sentence that they hear matches with the corresponding picture. The second one is similar but only one colorful picture is provided. In the third part, eight pictures with letters in alphabetic order (A-H) are shown. Test takers need to choose seven pictures in correct order based on sentences they hear. The last part requires test takers to select correct answers for questions according to the conversations they hear. The source of listening tasks is previously audio recorded.

The reading section has three parts with 30 test items. Part one is also yes or no. Test takers are asked to judge whether sentences they have read match the pictures. Part two is a cloze of five MC questions with five blanks in a short text. The final part is reading comprehension. Test takers are required to answer questions after reading a poster and a report.

Table 2. Descriptions of the GEPT-Kids Administered During March to May, 2017.

Section and part	Item	Score	CEFR Level
Listening			
Yes or no	1-5	5	A1
	6-11	6	A1
Ordering	12-18	7	A1
Conversation	19-25	7	A1
Reading			
Yes or no	1-20	20	A1
Cloze	21-25	5	A2
Reading comprehension	26-30	5	A2
Total		55	

3.2.2. DIF Questionnaire

In terms of content analysis through three reviewers, the DIF Questionnaire adapted from the

Geranpayeh and Kunnan (2007) was employed in the current study. This 5-point Likert scale questionnaire was used for reviewers to rate the suitability of first the test items that were flagged by the statistical procedure and the rest without displaying DIF. The scale was from 1 (strongly advantage) to 2 (advantage) to 3 (neither advantage nor disadvantage) to 4 (disadvantage) to 5 (strongly disadvantage). Below each rating was a blank for the reviewers to fill in their further explanations of judgements and comments on items.

3.3. Data Analysis

A mixed-methods sequential explanatory design was used in the current study with two phases (Creswell et al., 2003; Ivankova et al., 2006). In phase I, statistical analyses were conducted to detect item DIF. In phase II, content analyses attempted to identify the source of DIF and identify potential items with DIF. In other words, quantitative data are collected and analyzed first and then qualitative data are used for further explanation and elaboration (Ivankova et al., 2006).

3.3.1. Statistical Analysis

3.3.1.1. Test Performance Analysis

The Statistical Package for the Social Sciences (SPSS 24.0) was employed for data analysis. Normality was checked first with the result that the skewness (-1.00) and kurtosis (-.29) are within ± 2, indicating a reasonably normal distribution (Bachman, 2004). Descriptive analyses were conducted to calculate test mean scores of males and females including total scores and scores of each task and the spread and distribution of the scores. Item analysis was conducted to check item difficulty and item discrimination. An independent-samples *t*-test was conducted to compare test performances from each task to total scores between males and females.

3.3.1.2. Mantel Haenszel DIF Analysis

Test performance data were calculated via a non-IRT method, the Mantel-Haenszel (MH) method (Dorans & Holland, 1993; Holland & Thayer, 1988; Mantel & Haenszel, 1959). The MH method procedure was first used for matched groups (Mantel & Haenszel, 1959) and later adapted for detecting DIF by Holland and Thayer (1988). This is a classical approach and is a widely used method for DIF detection (e.g., Ryan & Bachman, 1992; Wu, 2009). MH analysis is prevalent because of its ease of computation and implementation and capacity of handling small sample sizes (Amirian et al., 2014; Fidalog et al., 2014; Padilla, 2012; Rogers & Swaminathan, 1993). In addition, it is a method for examining binary items. In the present study, the GEPT-Kids test items were all dichotomously scored. The items were graded as correct (1 point) or incorrect (0 point). All the test items were scored by three reviewers and the inter-rater reliability was checked.

The Mantel-Haenszel statistics were calculated with DIFAS 5.0 (Penfield, 2012). The DIFAS is designed for conducting common nonparametric DIF detection procedure (Penfield, 2005). The software is user-friendly and free of charge (Sirikit et al, 2016; Van den Broeck et al., 2013) and it can be used for both dichotomously scored items and polytomously scored items. For dichotomously scored items, the indices for DIF procedure contain Mantel-Haenszel Chi-Square (MH CHI), Mantel-Haenszel Common Log-odds Ratio (MH LOR), Standard Error Mantel-Haenszel Common Log-odds Ratio (LOR SE), Standardized Mantel-Haenszel Log-Odds Ratio (LOR Z), Breslow-Day (BD) Chi-Square, Combined Decision Rule (CDR), and ETS classification scheme (ETS) (Penfield, 2012).

The interpretations of indices for detecting DIF by DIFAS 5.0 are as below. The Mantel chi-square statistic is distributed as a chi-square with one degree of freedom (Penfield, 2012). The

higher the MH CHI value, the higher probability of the test item demonstrates DIF. As for LOR Z value, LOR Z value＞ 2or LOR Z value ＜ -2 shows DIF exists while -2 ≤ LOR Z ≤ 2 indicates no DIF is presented. If items show evidence of DIF, attention should be paid to MH LOR, from which it can be seen items with DIF showing favor to a reference group or focal group. Positive values indicate DIF favor a reference group. Negative values indicate DIF favor a focal group. BD is also distributed as a chi-square with one degree of freedom but it has been proved that it can detect non-uniform DIF. Non-uniform DIF affects students in specific ranges of the total score inconsistently (Martinková et al., 2017). The higher value shows higher probability of DIF. According to Penfield (2003), if an item shows either significant MH CHI or BD at a Type I error rate of .025, CDR would flag that item, displaying FLAG in the software. Otherwise, OK is displayed. Finally, ETS classification scheme established by ETS sets three categories for DIF with different degrees. Category A contains items with small DIF that can be considered as items functioning properly. Category B includes items with small to moderate DIF, which means that the items might be used, and Category C serious or large DIF, which are not preferred only if there is no other choice meeting test specification (Ryan & Bachman, 1992; Zikey, 2003).

3.3.2. Content Analysis

As Ferne and Rupp argued (2007), merely reporting the existence of DIF may not be enough and more attention should be given to identifying the sources of DIF by using a posteriori item content analysis. In the present study, content analysis was conducted to collect reviewers' opinions on the source of gender DIF flagged in the statistical procedure and the existence of potential bias towards gender in the GEPT-Kids. Selected reviewers examined items flagged with gender DIF by filling the DIF questionnaire with 5-point scale from strongly disadvantage to strongly advantage, which helps gain deeper and more comprehensive understanding of test reviewers' perceptions on sources of DIF. They added their explanations and comments on their judgements. In addition, they were asked to review the rest of the test items without DIF to spot items that might be in favor of males or females. The additional analysis is able to widen the scope of the content review analysis (Geranpayeh & Kunnan, 2007).

4. Results and Discussion

4.1. Test Performance Results

Table 3 presents the descriptive statistics including mean and standard deviations by two gender groups for the GEPT-kids, enumerating the test scores of each task of the Listening and Reading, Listening (LSUM), Reading (RSUM), sum scores of the whole test and the distribution of scores. Females had a slightly better performance than males in GEPT-Kids. Considering the mean scores for each task, males had lower values than females in all cases.

An analysis of each item in terms of item difficulty and item discrimination was conducted. Referring to item difficulty, the value of 50 items is above .70, indicating the items were easy for test takers. As for item discrimination, the value of eight items is lower than .30, which could be considered as marginal items in need of revision (Ebel & Frisbie, 1986). Cronbach's Alpha for the whole test is .93, showing high internal consistency reliability.

Table 3. Test Performances of Males and Females in GEPT-Kids (N=492).

	Mean	SD	Cronbach's Alpha
Listening			.88
Yes or no			

M	4.41	.86	.39
F	4.54	.75	
Yes or no			.64
M	4.98	1.32	
F	5.28	1.13	
Ordering			.76
M	6.24	1.38	
F	6.59	1.00	
Conversation			.79
M	4.84	2.08	
F	5.28	1.93	
LSUM			.88
M	20.47	4.07	
F	21.70	3.97	
Reading			.88
Yes or no			.77
M	16.95	3.03	
F	17.66	2.67	
Cloze			.71
M	3.82	1.50	
F	4.19	1.17	
Reading comprehension			.74
M	3.35	1.62	
F	3.74	1.52	
RSUM			.88
M	24.12	5.52	
F	25.59	4.71	
Total			.93
M	44.59	9.70	
F	47.29	8.34	

Note. LSUM=Listening sum; RSUM=Reading sum

4.2. Mantel-Haenszel DIF Results

Table 4 illustrates three test items that are worth discussing. The positive values of MH common log-odds ratio (MH-LOR) states that the two items favored males test takers. According to the index shown in DIFAS, critical values of Breslow-Day (BD) Chi-Square are able to indicate non-uniform DIF. Unlike the two items, the negative value of MH LOR demonstrated that L13 had advantages for females. With respect to the ETS classification scheme, two items were flagged as B. The listening item and the reading items were easier for males after matching ability. Both the items with DIF were in yes or no questions, where test takers were asked to judge whether what they hear or read are in accordance with the pictures shown in the test booklet.

To sum up, it was observed that two items manifest moderate DIF in the test from the MH analysis. They were from different sections. Meantime, one item displayed non-uniform DIF in listening. Little evidence from the MH DIF analysis indicates that males and females have different reactions to items in the GEPT-Kids.

Table 4. MH DIF Results in Terms of Gender by DIFAS.

Item	Favored	MH CHI	LOR Z	MH LOR	BD	CDR	ETS
L2	M	4.41	2.22	.78	.38	OK	B
L13	F	1.64	-1.43	-.78	5.79	FLAG	A
R2	M	4.94	2.29	.48	.04	OK	B

Note. P< .05

4.3. Content Analysis

Referring to content analysis, two items exhibited uniform DIF favoring males and one item exhibited non-uniform DIF favoring females were subjected to the review from three reviewers. They did not indicate L2 and L13 advantage or disadvantage either gender groups based on their ratings and comments, even the statistics showed evidence that these two items were with gender DIF in favor of males and females respectively.

4.3.1. Item L2

Item L2 (Mean=3.0) was considered as showing neither advantage nor disadvantage towards the two gender groups. The subskill of L2 is test takers' ability to understand vocabulary in common use and simple sentences. One of the reviewers commented that socks were supposed to be objects used in everyday life and both males and females should be familiar. Thus, the reviewer did not argue that the test item favors males or disadvantage towards females. As for the flagged DIF item, one of the reviewers speculated that males and females might have different preferences for colors or they might be sensitive to a certain type of color. In the picture of L2, there is a pair of green socks. Test takers' reactions towards the color of the picture might lead to the DIF item.

4.3.2. Item L13

Similarly, none of the test reviewers rated L13 (Mean=3.0) as biased. The primary target of the item is to examine whether test takers can match the picture with the description they hear. One reviewer left a comment that brushing teeth, which was clearly demonstrated in the picture, was a daily routine for both male and female test takers. Thus, the item was rated as neither disadvantage nor advantage. There were two key phrases "go to sleep" and "brushing his teeth" in the descriptions. Test takers might be more familiar with the former phrase because the word "brush" might be more difficult for them. Although L13 exhibited non-uniform DIF, possibly, the real difference in English language abilities between two gender groups might be the main reason to cause DIF.

4.3.3. Item R2

R2 was of high difficulty (65% correct) with quite poor discrimination (.09), which indicates that the item was not easy for test takers. Ratings from reviewers were entirely opposed to the information gathered from statistics. Reviewer A and B both rated 4.0 for males and 2.0 for females. Both of them deemed that the item was biased against male test takers. While reviewer C disagreed with it. R2 requires test takers to read the sentence and judge whether it is in line with the picture provided. The subskill is test takers' ability to understand vocabulary in daily use and sentences that consisted of simple words. The sentence in R2 was "Nancy is wearing glasses and pants". The picture provided is a female nurse wearing glasses and a nurse cap with a document in hand. The correct answer is no because the female in the picture wears a dress. Reviewer A commented that male and female test takers usually had different dress styles and females might be more sensitive to the dress demonstrated in the picture. Reviewer B left a similar comment that male test takers might not pay enough attention to dressing as females. Nevertheless, the item was detected in favor of males instead of females. Reviewers' views on items were not in line with results from statistics.

Besides the three flagged items, reviewers also conducted additional analysis on other items. Three reviewers came to consensus that 51 test items had a rating of 3.0 (neither advantage nor

disadvantage), showing a majority of test items were free of gender bias. However, four items in reading were rated above or below 3.0 (see Table 5).

Table 5. Items Rated above or below 3.0.

Items	Males			Mean	Females			Mean
	A	B	C		A	B	C	
R2*	4.0	4.0	3.0	3.7	2.0	2.0	3.0	2.3
R28	3.0	3.0	2.0	2.7	3.0	4.0	3.0	3.3
R29	3.0	3.0	2.0	2.7	3.0	4.0	3.0	3.3
R30	3.0	3.0	2.0	2.7	3.0	3.0	3.0	3.0

Note. A=Reviewer A; B=Reviewer B; C=Reviewer C.
1=strongly advantage; 2=advantage; 3=neither advantage nor disadvantage; 4=disadvantage; 5=strongly disadvantage.
*Items displayed DIF by statistical analysis.

4.3.4. Items R28, R29 and R30

Items R28, R29 and R30 are embedded in reading comprehension. They are from the same reading passage *Knowing the World: Animal* discussing snakes. For males, the three items were all averagely rated 2.7, indicating that the items have slight advantage for males. The three items received ratings that indicated bias towards female test takers were not supported by DIF statistics. R28 and R29 were rated 3.0 for males while 4.0 for females by reviewer B. As for reviewer C, R28 and R29 were rated 2.0 for males while 3.0 for females. In fact, they conveyed the same idea that R28 and R29 advantaged male test takers. The primary subskill of R28 is the ability to identify the main idea of paragraphs. For R29, the target is to test the ability to recall details from the paragraph. Similarly, the two reviewers concluded that it was the topic of the reading passage that led to the bias towards females. They had similar comments that males might have slight advantage because more males were into bugs or reptiles than females. Reviewer B mentioned that snakes might cause uncomfortable feelings for females because many of them were afraid of snakes, which might account for the disadvantage for them while they were taking the test. As for R30, reviewer C left a similar view that the topic led to the rating showing preference for males.

4.4. Limitations and Suggestions for Further Research

An effort was made to detect gender DIF in the GEPT-Kids and provide explanations for the causes in the present study. But some limitations existed in the whole research. First, the sample size of the research was small, which restricted the choices of DIF analysis method, e.g., three-parameter unidimensional IRT model. Only the MH method was utilized, which is undesirable for it cannot ensure the certainty of numbers of DIF. Relatively large sample size is required by a number of more sophisticated statistical methods (Ferne & Rupp, 2007). Moreover, MH method can only detect uniform DIF. Although DIFAS can flag non-uniform DIF, it failed to convey details regarding which group was favored at which ability level. Thus, further research of the items showing non-uniform DIF cannot be continued. Second, except for listening and reading, writing is one of the parts in the regular GEPT-Kids. Due to insufficient time for test administration, the writing was eliminated during data collection. Third, only items in the multiple-choice format were included in the research. Whether male or female test takers have advantage on different test formats is another question under-researched and the answer remains unknown. Therefore, further research may

consider taking the writing section into consideration and laying emphasis on different response formats.

5. Conclusion

The present study investigated gender-based DIF in the GEPT-Kids administered in five Chinese speaking cities. A slight difference in test performance was found between two gender groups and females had better performance in all cases. Two phases were conducted in the process of DIF analysis. First, test items were analyzed through statistical methods. Secondly, items that demonstrated DIF were subject to reviewers. Three items were examined with DIF by using MH analysis method on DIFAS 5.0 for its computation simplicity and capability for DIF detection in small sample sizes. Two items exhibited moderate DIF, both of which showed preference for males. One item displayed non-uniform DIF favoring females. All the items were also subject to reviewers for content analysis. Totally, four items were identified with possibility that advantaged either gender group. But there was no clear evidence could explain the probable causes for the gender DIF or potentially biased items from observations to content reviews. Item content and display were mentioned as the most possible causes. It seems that it is the discrepancy in language ability between two gender groups accounting for the existence of DIF. Thus, it can be concluded that the GEPT-Kids is possibly not biased against either males or females.

Although no systematic DIF was detected in the GEPT-Kids, the study has some implications for test developers and future researchers. Even though GEPT-Kids is not a high-stakes assessment, DIF analysis is still necessary for further improvement of the test quality and assurance of test fairness. For test item outlining, test developers should be aware that test content might advantage or disadvantage a particular group. Therefore, test developers should be careful in choosing topics and materials to ensure the authenticity of the test. As a test designed for primary school students, test items illustrated with pictures can be attractive and motivate young English learners to complete the test. But more consideration should be taken into account in terms of the design of pictures illustrated beside test items.

Acknowledgements

First and foremost, we would like to extend our sincere gratitude to Prof. Antony John Kunnan for his useful suggestions, incisive comments and constructive criticism. Moreover, our special thanks go to the Language Training and Testing Center (LTTC, Taiwan) for granting the data. Any progress that we have made is the result of their profound concern and selfless devotion. Furthermore, we really to want to say thanks to our language assessment seminar research (LASeR) group members from whom we have learnt so much.

References

[1] Ahmadi, A., & Bazvand, A. D. (2016). Gender Differential Item Functioning on a National Field-Specific Test: The Case of PhD Entrance Exam of TEFL in Iran. Iranian Journal of Language Teaching Research, 4(1), 63- 82.

[2] Alderman, D. L., & Holland, P. W. (1981). Item performance across native language groups on the Test of Egnlish as a Foreign Language. ETS Research Report Series.

[3] American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). Standards for educational and psychological testing. AREA.

[4] Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting gender DIF with an English proficiency test in EFL context. Iranian Journal of Language Testing, 4(2), 187-203.

[5] Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-23). Lawrence Erlbaum Associates, Inc.

[6] Aryadoust, V., Goh, C. C., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB

listening test. Language Assessment Quarterly, 8(4), 361-385.
[7] Bachman, L. F. (2004). Statistical analyses for language assessment. Cambridge University Press.
[8] Bailey, A. L. (2008). Assessing the language of young learners. In N. H. Hornberger (Ed.), Encyclopedia of language and education (pp. 379-398). SpringerLink. https://doi.org/10.1007/978-0-387-30424-3_188
[9] Banerjee, J., & Papageorgiou, S. (2016). What's in a Topic? Exploring the Interaction Between Test-taker Age and Item Content in High-Stakes Testing. International Journal of Listening, 30(1-2), 8-24.
[10] Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27(2), 165-174.
[11] Boyle, J. P. (1987). Sex differences in listening vocabulary. Language Learning, 37(2), 273-284.
[12] Brantmeier, C. (2003). Beyond linguistic knowledge: Individual differences in second language reading. Foreign Language Annals, 36 (1), 33-43.
[13] Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items (Vol. 4). Sage.
[14] Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. ETS Research Report Series.
[15] Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. Language Testing, 2(2), 155-163.
[16] Chiu, H. H. (2008). Are there any gender differences in the GEPT picture description listening comprehension test? Chia Nan Annual Bulletin: Humanity, 34, 409-422.
[17] Conoley, C. A. (2003). Differential item functioning in the Peabody Picture Vocabulary Test (3rd Edition): Partial correlation versus Expert judgment. PhD Thesis, Texas A&M University.
[18] Creswell, J. W., Tashakkori, A., Jensen, K. D., & Shapley, K. L. (2003). Teaching mixed methods research: Practices, dilemmas, and challenges. In A. Tashakkori, & C. Teddlie (Eds.), Handbook of mixed methods in social and behavioral research (pp. 91-110). Sage Publications.
[19] DeMars, C. E. (2000). Test stakes and item format interactions. Applied Measurement in Education, 13(1), 55-77.
[20] Dorans, N. J., & Paul, W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W.Holland & H. Wainer (Eds.), Differential item functioning (pp.3-23). Lawrence Erlbaum Associates, Inc.
[21] Ebel, R. L. & Frisbie, D. A. (1986). Essentials of educational measurement (4th Ed.). Prentice-Hall.
[22] Ellis, L., & Ficek, C. (2001). Color preferences according to gender and sexual orientation. Personality and Individual Differences, 31(8), 1375-1379.
[23] Fernandes, A. C. (2015). Gender differential item functioning on English as a foreign language pragmatic competence test: Implications for English assessment policy in China. PhD Thesis, Niagara University.
[24] Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. Language Assessment Quarterly, 4(2), 113-148.
[25] Fidalgo, Á. M., Hashimoto, K., Bartram, D., & Muñiz, J. (2007). Empirical bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. The Journal of Experimental Education, 75(4), 293-314.
[26] Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. Language Assessment Quarterly, 4(2), 190-222.
[27] Hasselgreen, A., & Caudwell, G. (2016). Assessing the language of young learners. Equinox Publishing.
[28] Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. Test Validity, 129-145.
[29] Hoyenga, K. B., & Wallace, B. (1979). Sex differences in the perception of autokinetic movement of an afterimage. The Journal of General Psychology, 100(1), 93-101.
[30] Ivankova, N. V., Creswell, J. W., & Stick, S. (2006). Using mixed-methods sequential explanatory design: From theory to practice. Field Methods, 18(3), 3-20. https://doi.org/10.1177/1525822X05282260
[31] Jiao, H., & Chen, Y. (2014) Differential item and testlet functioning analysis. In A. J. Kunnan (Ed., 1st volume), The Companion to Language Assessment (pp.1282-1300). John Wiley & Sons.
[32] Kunnan, A. J., (1990). DIF in native language and gender groups in an ESL placement test. TESOL Quarterly, 24(4), 741-746.
[33] Kunnan, A. J. (2004). Test fairness. In M. Milanovic, & C. Weir (Eds.), Europe language testing in a global context: Selected papers from the ALTE conference in Barcelona (pp.27-48). Cambridge University Press.
[34] Kunnan, A. J. (2007). Test fairness, test bias, and DIF. Language Assessment Quarterly, 42(2), 109-112.
[35] Kunnan, A. J. (2010). Fairness matters and Toulmin's argument structures. Language Testing, 24(2), 183-189.
[36] Kunnan, A.J. (2017). Evaluating language assessments. Routledge.
[37] Lei, X. (2007). Shanghai gaokaoyingyufenshu de xingbiechayi he yuanying [Gender differences and their sources on the National Maculation English Test in the Shanghai area]. Shanghai Research on Education, 6, 43-46.
[38] Liao, Y. (2016). Gender differences and differential item functioning on the English GSAT multiple-choice questions. Soochow Journal of Foreign Languages and Cultures, (41), 21-59.

[39] Liu, B. & Li, Y. (2010). Opportunities and barriers: Gendered reality in Chinese higher education. Frontiers of Education in China, 5, 197-221.
[40] Lin, J., & Wu, F. (2004). Differential performance by gender in foreign language testing. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, USA.
[41] Lorenz, R. (2016). Does gender make a difference? Gender-related fairness of high-stakes testing in A-level examinations in English as foreign language in the German state of North Rhine-Westphalia in the context of Educational Governance. Journal for Educational Research Online, 8(2), 10-30.
[42] Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. CBE-Life Sciences Education, 16(2), 2.
[43] Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719-748.
[44] McKay, P. (2006). Assessing young language learners. Cambridge University Press.
[45] Mullis, I. V., Martin, M. O., Foy, P. & Drucker, K. T. (2012). PIRLS 2011 International Results in Reading. International Association for the Evaluation of Educational Achievement, Amsterdam, the Netherlands.
[46] O'Neill, K. A., McPeek, W. M., (1993). Item and test characteristics that are associated differentialitem functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-23). Lawrence Erlbaum Associates, Inc.
[47] O'Neill, K. A., McPeek, W. M., & Wild, C. L. (1993). Differential Item Functioning on the Graduate Management Admission Test (ETS-RR-35). Educational Testing Service.
[48] Padilla, J. L., Hidalgo, M., Benítez, I., & Gómez -Benito, J. (2012). Comparison of three software programs for evaluating DIF by means of the Mantel-Haenszel procedure: EASY-DIF, DIFAS and EZDIF. Psicológica, 33(1), 135-136.
[49] Pae, T. I. (2004). DIF for examinees with different academic backgrounds. Language Testing, 21(1), 53-73.
[50] Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. Language Testing, 29(4), 533-554.
[51] Papp, S., & Rixon, S. (2018). Examining young learners: Research and practice in assessing the English of School-age Learners. In N. Saville, & C. J. Weir (Eds., Vol. 47), Studies in language testing. Cambridge University Press.
[52] Park, G. P. (2008). Differential item functioning on an English listening test across gender. TESOL Quarterly, 42(1), 115-123.
[53] Penfield, R. D. (2003). Applying the Breslow-Day Test of Trend in Odds Ratio Heterogeneity to the Analysis of Non-uniform DIP. Alberta Journal of Educational Research, 49(3), 231-243.
[54] Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. Applied Psychological Measurement, 29(2), 150-151.
[55] Penfield, R. D. (2012). DIFAS: 5.0 user's manual. http://erm.uncg.edu/wp-content/uploads/2012/07/ DIFASManual_V5.pdf
[56] Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. Language Testing, 9(1), 12-29.
[57] Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17(2), 105-116.
[58] Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. Journal of Educational Measurement, 27(2), 109-131.
[59] Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6(4), 317-375.
[60] Shohamy, E. (1984) Does the testing method make a difference? The case of reading comprehension. Language Testing, 1, 147-170.
[61] Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. Language Testing, 8, 23-40.
[62] Sirikit, R., Choptham, M., Mahalawalert, P., Tagontong, N., & Apinyapibal, S. (2016). An investigation of differential item functioning and differential test functioning of SWUSAT during 2010-2013. Scholar: Human Sciences, 8(2).
[63] Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. Papers in Language Testing and Assessment 4 (1), 97-124.
[64] Swinton, S. S., & Powers, D.E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups. ETS Research Report Series.
[65] Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. Language Testing, 17(3), 323-340.

[66] Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. Language Testing, 22(2), 211-234.
[67] Van den Broeck, J., Bastiaansen, L., Rossi, G., Dierckx, E., & De Clercq, B. (2013). Age-neutrality of the trait facets proposed for personality disorders in DSM-5: A DIFAS analysis of the PID-5. Journal of Psychopathology and Behavioral Assessment, 35(4), 487-494.
[68] Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. Journal of Psychoeducational Assessment, 29(4), 364-376.
[69] Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions, Educational Assessment, 10(1), 1-17.
[70] Wu, J. (2009). Differential Item Functioning in gender and living background groups in the GEPT. Paper presented at the 13th International Conference on Language Education, Kaohsiung, Taiwan.
[71] Zhu, Y. Y., Wu, Q. M., & Jiao, L. Y. (2017). Analysis of Differential Item Functioning in A Primary English Test. China Examinations, 4, 54-56.
[72] Zieky, M. (2003). A DIF primer. http://www.ets.org/Media/Tests/PRAXIS/pdf/DIF_primer.pdf
[73] Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, 4(2), 223-233.