Gender Differential Item Functioning Analysis of the University of Tehran English Proficiency Test

Abbas Ali Rezaee*

Assistant professor, Faculty of Foreign Languages and Literatures, University of Tehran, I.R. Iran

Enayatollah Shabani**

PhD student of TEFL, Faculty of Foreign Languages and Literatures, University of Tehran, I.R. Iran

(Received: 8 Dec. 2009, Accepted: 4 Feb. 2010)

Abstract

The University of Tehran English Proficiency Test (UTEPT) is a high-stakes entrance examination taken by more than 10,000 master's degree holders annually. The examinees' scores have a significant influence on the final decisions concerning admission to the University of Tehran Ph.D. programs. As a test validation investigation, the present study, which is a bias detection research in nature, utilized multistep logistic regression (LR) procedure to examine the presence of gender differential item functioning (DIF) in the UTEPT with a sample of 6,555 examinees who took the test in November 2006. Specifically, the LR DIF two-degree of freedom Chi-squared test of significance was employed to test the significance of DIF. Following what has been recently suggested in the literature, the test of significance for DIF was accompanied by a measure of magnitude, namely the *R*-squared effect size for LR DIF, for the interpretation of which the two widely accepted classification schemes were used comparatively. The results reveal that 39 of the 100 items in the test display significant gender differences. However, these group differences are viewed as "negligible" based on both of the schemes. Accordingly, it could be argued that the UTEPT is a gender DIF-free test, though the Reading Comprehension section of the test remains in need of further analysis as it seems that the general trend of DIF indices at the item level may hint at an inclination towards males.

Keywords: Differential Item Functioning, Logistic Regression, UTEPT, English Proficiency Test, Gender Differences, High-stakes Test.

---

* Tel: 021-61119081, Fax: 021-88634500, E-mail: aarezaee@ut.ac.ir

** Tel: 021-61119081, Fax: 021-88634500, E-mail: shabani@ut.ac.ir

I.Introduction

One of the important factors which should be taken into account in dealing with the validity of any test is the issue of fairness (Thissen, 2001). Educational Testing Service Fairness Review Guidelines (2003) offers a simple and fairly straightforward verdict: "A test that shows valid differences is fair; a test that shows invalid differences is not fair" (p. 2). Therefore, when people with similar abilities in the construct being measured perform substantially differently on a test item it is necessary that the item be reviewed for fairness (Gierl, Khaliq & Boughton, 1999), and perhaps removed if the differential performance of the examinees is not balanced or cancelled over the test as a whole (Wainer, Sireci & Thissen, 1991). This issue has made a great area of investigation for researchers, and a technique which empirically measures the differential functioning of items for groups which are matched on the ability of interest has now become "the new standard in psychometric bias analysis" (Zumbo, 1999, p. 6) and is now "a key component of validity studies in virtually all large-scale assessments" (Penfield & Camilli, 2007, p. 125). This technique is called differential item functioning (DIF) after Holland and Thayer who used the term in a seminal chapter on test validity (1988).

DIF occurs when the responses of individuals having the same ability of interest show systematic differences simply based on their membership in a certain group. It is necessary, therefore, that judgmental and statistical analyses be applied to detect the items in a test that function differentially for individuals with the same ability from different groups. It can be argued that the higher the stakes of a test, the more the need for such analyses. The necessity of conducting a study examining the issue of fairness in the University of Tehran English Proficiency Test (UTEPT) is best understood when one takes into account the fact that every year the test is taken by more than 10,000 masters degree holders and the inferences made on the basis of the test scores are a crucial factor in determining the admission of the test takers in the Ph.D. program at the University of Tehran.

Gender DIF analysis is an important part of test development as it helps to examine and eliminate the items which may be potentially unfair to some groups of

test takers because of gender group membership. To the knowledge of the present researchers, no such study has ever been done in Iran. The purpose of the present study, therefore, is to investigate the presence, the magnitude as well as the direction of DIF in the items of the UTEPT.

II. Background

DIF exists when after controlling for overall ability examinees from different groups have a different probability or likelihood of successfully answering an item. Though the development of statistical methods for identifying potentially biased items began in the 1970s (Penfield & Camilli, 2007), it seems that the first bias studies were done due to the Civil Rights Act of 1964 in the U.S. (Conoley, 2003 and Duncan, 2006). Owing to the Act, the first DIF studies were those which were conducted to ensure test fairness for black and white examinees. The methods used in the early DIF studies were ANOVA, correlational methods and Transformed Item Difficulty (TID) method which looked for differential difficulty as an indicator of bias. Since that time, many approaches have been developed to examine DIF, among which standardization (Dorans & Kulick, 1986), Mantel-Haenszel (MH; Holland & Thayer, 1988), logistic regression (LR; Swaminathan & Rogers, 1990), Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993) and item response theory (IRT) procedures are the most widely used statistical methods for detecting DIF in dichotomously scored responses (see Sireci & Allalouf, 2003 for a detailed list of DIF methods and the types of data for which each method is appropriate).

Generally, DIF is of two types, uniform and non-uniform. Uniform or unidirectional DIF exists when the probability of endorsing an item (answering an item correctly) is greater for one group than for the other group over *all* the levels of proficiency. The existence of non-uniform or crossing DIF demonstrates that the difference in probabilities of a correct response is not the same at all levels of proficiency between the two comparison groups. That is, the probability of correctly answering an item is higher for one group at some points on the scale, and higher for the other group at other points (see Salkind, 2007). In IRT terminology, non-uniform

DIF exists when the item response curve for the focal group intersects the item response curve for the reference group. If the item response curves for the focal and reference groups do not cross, it is said that DIF is uniform. In other words, uniform DIF occurs when there is no interaction between the ability level and group membership. If there is an interaction between ability and gender, this signals the existence of non-uniform DIF (Swaminathan & Rogers, 1990).

For many years, the prevalence of non-uniform DIF was an intense debate among researchers. This explains the introduction of a DIF examination procedure for both uniform and non-uniform DIF as late as 1990, that is, almost three decades after the first statistical methods for DIF detection were proposed. Though the controversy seems not to have ended up in a unanimous agreement, there is little doubt today as to the occurrence of non-uniform DIF (Jodoin & Gierl, 2001). Therefore, utilizing an approach which can detect both uniform and non-uniform DIF seems superior for the researchers to enjoy a higher standard of practice. With this aim in view, the LR procedure was used as the DIF detection procedure in the present study.

III. Previous Gender DIF Studies in Language/Verbal Tests

Though the issue of bias has long been a concern of language testers, the systematic study of it can only be observed in the studies conducted after 1960. Yet, it was only as late as the 1980s that with the advent of more sophisticated approaches to DIF examination researchers could screen language tests for what could possibly lead to the unfair treatment of examinees of different gender. In one of these studies, Lawrence and Curley (1989) found that the items related to technical science content were more difficult for females. Curley and Schmitt (1993) found that of the 2,250 SAT-Verbal questions pretested in 1990, only 8.5 percent displayed moderate and large amounts of gender DIF. Buck, Kostin and Morgan (2002) found that item content can be associated with gender-based performance differences. In a study conducted by Ryan and Bachman (1992), it was found that males and females do not react differently at the item level in TOEFL and FCE. Lin

and Wu (2003) conducted a DIF study on the EPT (English Proficiency Test) in China, which was modeled after the TOEFL and found that there was not much gender DIF in the test. As they reported, of a total of 120 items, only two items were identified as exhibiting C-level or large DIF and in fact 89% of the test items exhibited "no" DIF whatever.

Conoley (2003) conducted a study to detect DIF in a measure of receptive vocabulary and found that expert judges could not accurately predict items that demonstrate DIF against one group, and that it was more plausible not to rely heavily on the expert's subjective speculations in bias detection studies. The important point to note here is that DIF is a test-dependent phenomenon. Differential performance of the members of different gender groups on a certain item in a certain test does not reveal anything about the existence or direction of DIF in other language tests. In order to examine DIF in a test, an objective DIF analysis must be done on exactly that very test, and the findings of other DIF studies do not indicate much about the presence or absence of DIF in the test. Taking this point into account, the present research was an attempt to investigate the presence, magnitude, and direction of gender DIF in the UTEPT.

IV. Method

1. Data Set

The data used in this study consisted of the test results from the UTEPT which was taken by master's degree holders of different majors seeking entry to the University of Tehran Ph.D. program in the year 2006. The analysis was performed on 6,555 cases, 4553 (69.5%) males and 2002 (30.5%) females. The male group was considered as the reference group and coded 1, and the female group was taken as the focal group of the study, given the dummy code 2. (Though McNamara and Roever, 2006, consider such categorizations conceptually problematic, it is common practice in DIF studies.)

Broer, Lee, Rizavi and Powers, (2005) presume a large sample size to be a

requirement for the LR procedure. Schmitt, Holland and Dorans (1992) also suggest that the largest possible number of examinees should be used to obtain stable DIF estimates and ensure sufficient power for DIF detection. Curley and Schmitt (1993) recommend that in order to come up with reliable results, large sample sizes be used, particularly for the focal group. Therefore, the sample size of the present study seems to enjoy a considerable degree of acceptability for the results to be reliably interpretable.

2. UTEPT

The University of Tehran English Proficiency Test (UTEPT) is a test of English proficiency developed in the Language Testing Center of the Faculty of Foreign Languages and Literature, University of Tehran. It functions as a screen test which must be taken by the M.A./M.S. holders of different majors to be allowed to sit the major-specific Ph.D. entrance examination. The test is taken by more than 10,000 master's holders annually and the final decisions concerning admission to the University of Tehran Ph.D. candidacy are influenced by the results of this test. The test calibrates items separately within subtests and creates its final scale as a weighted composite of the scales created within subtests. However, the scores for the separate subtests are available and can be used for research purposes.

The UTEPT is a set of 100 four-option multiple-choice items with a time limit of 100 minutes. It includes three parts: The Structure and Written Expression section which comprises 40 items, the Vocabulary section which consists of 25 items, and the Reading Comprehension section which has 35 items. Throughout the study, care was taken to consider the developers' concern over the confidentiality of the test.

The development process of the UTEPT includes three main steps. In the first step, the items are written. Then their usefulness is evaluated through different stages of review, and finally the final form of the test is created. There is yet another stage of internal review which finalizes the selection of test items and is applied after the test items are selected. Although biased items might be identified through the review stages of test development, statistical procedures can provide the test

developers and users with a robust niche objectively and empirically as regards the fairness of the test.

3. Procedure

An approach first proposed by Swaminathan and Rogers (1990), the LR approach to examining DIF in dichotomous items is based on the examination of regression in predicting item response from main effects of group and ability and their interaction (Kanjee, 2007). In other words, in binary LR, the item response, which in binary scoring format can have two values (correct or incorrect), will be the dependent variable, and the grouping variable (e.g., male vs. female), the ability variable and a group-by-ability interaction term will be taken as independent variables.

In the present study, the dichotomously scored item-level data based on 6,555 examinees was subjected to the multistep LR model which uses group membership (g) and ability-by-interaction (g*θ) term to estimate the probability that a randomly selected examinee with an ability of θ answers an item correctly. The LR equation can be shown as $Y = b_0 + b_1\theta + b_2 g + b_3\ \theta * g$, where θ is the ability level of the examinee (which is usually represented by the total scale score), the letter g denotes the group membership variable which can take the two codes specified for the reference and focal groups, b_0 is the intercept for a dichotomized variable, and the parameters b_1, b_2 and b_3 are the slope parameters which denote the weights for ability, group membership, and the interaction of the two, respectively (Zumbo, 1999). The asterisk represents interaction between variables. In fact, this equation is a linear equivalent of the original LR DIF formula proposed by Swaminathan and Rogers (1990). The null hypothesis in LR DIF analysis can be shown as $\mathbf{H_0}: b_2 = b_3 = 0$ H_0:b_2=b_3=0 for each item.

The LR method for binary scores employed in the present study was the three-step modeling process proposed by Zumbo (1999). In this method, in the first step the conditioning variable (θ) is entered into the regression equation. Conditioning is one of the requirements for every comparison between the groups. In the second

step, the group membership variable (g) is entered into the equation, and finally, in the last step, the score-by-group interaction term (θ*g) is entered. In sum, the whole process can be described as the following:

Step 1: enter the conditioning variable (θ)

Step 2: enter the group membership variable (g)

Step 3: enter the term for the interaction between the conditioning variable and group variable (θ*g)

In practical terms, by subtracting the Chi-squared value of the model with the ability variable from the Chi-squared value of the model with the interaction term and comparing the results with its distribution function with 2 degrees of freedom, the statistical tests for DIF can be computed (the two degrees of freedom is the result of comparing the model Chi-squared statistic at Step 3, which is three, and the model Chi-squared statistic at Step 1, which is one). An item is said to exhibit significant group differences when the *p*-value for the Chi-squared test of significance for that item is less than or equal to .01.

However, simulation studies have indicated that using the LR procedure without a measure of effect size could result in situations in which Type I error rate would be higher than expected (Jodoin & Gierl, 2001). In fact, an effect size measure is a descriptive statistic representing the "magnitude" of DIF. Since different studies have suggested that the power of statistical test is sensitive to the employed sample size (e.g. Jodoin & Gierl, 2001 and Zumbo, 1999), it seems essential that a measure of magnitude be used to enable the researcher to interpret DIF in terms of size.

The effect size for LR DIF is tested by the *R*-squared coefficient. *R* is a statistic that is used to look at the partial correlation between the dependent variable and each of the independent variables. *R* can range from -1 to +1. Within this framework, R^2 R^2 changes are used to represent the magnitude of DIF. R^2 R^2 is the result of the ratio of model fit Chi-squared to -2Log likelihood. In effect, the R^2 values corresponding to uniform and non-uniform DIF are associated with b_2g and b_3θ*g, respectively. In this procedure, uniform DIF can be measured by comparing the R^2 value of the ability-only regression equation (i.e., Step 1) with the R^2R^2

value after adding the group variable (i.e., Step 2). Non-uniform effect sizes can be calculated by comparing the value of R^2 in Step 3 and Step 2, that is, after and before entering the interaction term into the LR equation. It is the aggregation of uniform and non-uniform effects that makes the total DIF effect size.

Zumbo and Thomas (1996) offer a classification guideline for DIF effect sizes. Based on this guideline, DIF values can be classified as negligible (A-level), moderate (B-level) or large (C-level). This guideline can be summarized as the following:

1. Negligible or A-level DIF: $R^2 < 0.13$
2. Moderate or B-level DIF: $0.13 \leq R^2 < 0.26$
3. Large or C-level DIF: $0.26 \leq R^2$

Therefore, putting the two procedures together, as Zumbo (1999) points out, for an item to be classified as displaying DIF, the two-degree-of-freedom Chi-squared test in logistic regression had to have had a *p*-value less than or equal to 0.01 (set at this level because of the multiple hypotheses tested) *and* the Zumbo-Thomas effect size measure had to be at least an *R*-squared of 0.130. (p. 27) [italics original]

It should be noted, however, that this is not the only classification scheme for R^2 values in the DIF literature. In a simulation study on evaluating Type I error and power rates using an effect size with logistic regression DIF, Jodoin and Gierl (2001) compared the effect size measures for logistic regression and SIBTEST to suggest a different set of classification standards for R^2 changes, which they believed could be considered strikingly different form the one proposed by Zumbo and Thomas (1996). In brief, the guideline proposed by Jodoin and Gierl can be summarized as the following:

1. Negligible or A-level DIF: $R^2 < 0.035$
2. Moderate or B-level DIF: $0.035 \leq R^2 < 0.070$
3. Large or C-level DIF: $0.070 \leq R^2$

We used both of these classification schemes in this study.

4. Analysis

In this study the multistep LR procedure was used to examine the existence, the magnitude, and the direction of DIF in the items of the UTEPT. For this purpose, first the grouping variable needed to be coded and defined as nominal. Dummy codes1 and 2 were used for the reference and focal groups respectively. Also, the total score for each individual was needed to be specified to be used as the conditioning variable. The item responses were then coded dichotomously. That is, all the 100 item responses for the entire 6,555 examinees were converted to a 0-and-1 system of coding.

The dependent variable (item response) and the independent variables (ability and gender) were entered into SPSS and the binary logistic analysis was used to analyze the data. For each item, the simultaneous uniform and non-uniform DIF two-degree of freedom Chi-squared (DIF χ^2[2]) was produced, which was the test of significance for DIF in LR. The R^2 changes for the items with significant group differences were then estimated and compared against the Zumbo-Thomas (1996) and Jodoin-Gierl (2001) classification schemes to find the items with moderate or large DIF indices.

V. Results

Of the 6,555 test takers whose records have been used in the present study, 4553 were male and 2002 were female. The average total score of all the participants in the test was 54.71. The means of the total scores for males and females were found to be 54.12 and 56.04 respectively. A preliminary analysis of the data indicated that the standardized mean difference in total scores of males (M = -.036, SD = 1.01) and females (M = .083, SD = .96) was d = -.12, which was found to be statistically significant, t(6553) = -4.47, $p \leq 0.01$ (two-tailed). However, it would be viewed as a very small effect size using Cohen's (1988) standard of .20 as "small". Table 1 below gives a summary of the descriptive statistics.

Table 1. Descriptive statistics of the male and female test-takers' scores

	Number	Mean of Total Score	Standard Mean Difference	SD
Male	4553	54.12	-.036	1.01
Female	2002	56.04	.083	.96

The results of the formal test of significance for DIF revealed that of the 100 items in the test, 39 items could pass the $p \leq 0.01$ condition for the two-degree-of-freedom Chi-squared test. Table 2 presents results of the DIF analyses for these 39 items. The results indicated that non-uniform DIF was not prevalent in the UTEPT and from the 39 items which could pass the test of significance, 24 items were found not to exhibit any non-uniform DIF whatever. This means that 85 percent of the items in the entire test displayed no non-uniform DIF. On the whole, items 51 and 83 exhibited the largest gender effect sizes, R^2=0.01 (see Table 2).

Table 2. Uniform, Non-uniform, Total R2 Effect Sizes, and the Chi-squared test results for the items with p ≤ 0.01

Item No.	R^2 effect size			χ^2 test	
	Uniform	Non-uniform	Total	χ^2 (2)	*P*

Item No.	R^2 effect size			χ^2 test	
	Uniform	Non-uniform	Total	χ^2 (2)	*P*
2	.002		0.002	14.669	.001
4	.004		0.004	21.853	.000
6	.004		0.004	23.624	. 000
10	.005		0.005	28.749	. 000
12	.003		0.003	15.496	. 000
14	.004		0.004	22.437	. 000
17	.004		0.004	25.504	. 000
21	.002		0.002	13.879	. 001
26	.007	.002	0.009	55.320	. 000
27	.007		0.007	38.692	. 000
33	.002	.001	0.003	15.314	. 000
42	.002		0.002	9.151	. 010
43	.008		0.008	37.462	. 000
46	.005		0.005	24.302	. 000
47	.007		0.007	34.131	. 000
48	.007		0.007	44.839	. 000
49	.003		0.003	21.120	. 000
51	.010		0.010	63.276	. 000
57	.003	.001	0.004	27.817	. 000
58	.002	.002	0.004	22.071	. 000
59	.002	.001	0.003	16.771	. 000
61	.004		0.004	20.586	. 000
62	.003		0.003	16.439	. 000
64	.003	.001	0.004	23.636	. 000
68	.003		0.003	15.780	. 000
70	.002	.001	0.003	13.368	. 001
74	.007	.001	0.008	46.856	. 000
75	.001	.002	0.003	15.213	. 000
77	.004		0.004	19.407	. 000
79	.002	.001	0.003	14.224	. 001
82	.002	.002	0.004	21.047	. 000
83	.008	.002	0.010	56.992	. 000
85	.005		0.005	27.947	. 000
91	.005	.001	0.006	31.481	. 000
93	.003		0.003	14.695	. 001
94	.007		0.007	33.200	. 000
95	.006		0.006	42.248	. 000
99	.002	.001	0.003	18.410	. 000
100	.003	.001	0.004	20.727	. 000

Note. * $p \leq 0.01$ two-tailed.

VI. Discussion and Conclusions

The present study was an attempt to examine the fairness of the UTEPT with regard to gender. It can offer some insights into the test design and development process at the University of Tehran. Although there are some researches to suggest that multiple choice questions might favor either boys or girls (Woods, 1991), the general consensus is that at the item level, no group differences should be found in a test of English language proficiency. Gender group membership should not have any impact on the performance of examinees on an item. This is of great prominence

since the UTEPT is a high-stakes test with a major impact on people's future academic and social life. Though the existence of DIF in a test does not necessarily signal bias, the present study was conducted as a prelude to more profound analyses towards developing unbiased tests.

The results revealed that of the whole 100 items in the test, 61 items could not pass the formal test of significance for DIF. That is, the two-degree of freedom Chi-squared test for these items did not have a *p*-value less than or equal to .01. This left 39 items (39% of the test items) to be tested for the magnitude of DIF. Once the effect sizes for the items with significant gender differences were estimated, these values needed to be interpreted based on the classification schemes for DIF. If an item with significant group differences exhibited *moderate* or *large* DIF, it could be considered as a potential threat to the fairness of the test. In other words, it can be suggested that these test items should be replaced with ones which show less DIF.

Based on both Zumbo-Thomas (1996) and Jodoin-Gierl (2001) guidelines, all of the 39 items with significant gender differences were found to display negligible or A-level DIF. In consequence, it can be concluded that concerning gender the UTEPT appears to be a DIF-free test. This, however, can only be regarded as a general conclusion since it only takes the aggregate results into account. It is not implausible to argue that 39 out of 100 is still a lot, though the magnitudes for DIF indices were shown not to have reached the acceptable level to be considered moderate or large. What is of equal importance in interpreting the results is the individual direction of DIF indices and the cumulative effect of these indices. It seems necessary therefore to examine whether these "negligible" DIF values display random preferential treatment for either of the gender groups or it is the case that a considerable number of the indices favor one group more than the other, in which case, the argument supporting that the test is gender DIF-free can hardly remain persuasive.

The individual analyses of the 39 items which could pass the test of significance for DIF displayed that 25 items favor males and 14 items favor females. This suggests that on the whole these "negligible" DIF indices are shown to have favored

males more than females. It is even possible to have a closer look at the direction of the DIF values by considering the three parts of the test and comparing the direction of the items that favor males and those that favor females in the three sections, namely Structure and Written Expression, Vocabulary, and Reading Comprehension.

Out of 40 items in the Structure and Written Expression Structure section, 11 items contained DIF values which could pass the formal test of significance. From these 11 items, six favored males and five favored females. Taking the Vocabulary section into account, it was found that out of the 25 items in this section, 13 items displayed significant DIF values, eight favoring males, and five favoring females. Finally, in the Reading comprehension section, 15 out of 35 items in the section were found to have significant DIF indices, out of which 11 favored males and four favored females.

Putting the magnitude of the found DIF values aside for the moment, this subsequent exploratory analysis of the results reveals that with regard to the differential performance of males and females on different sections of the test, it is the Reading Comprehension section of the test which for the most part appears to be responsible for the differential performance of males and females on the items of the UTEPT (note again that the group differences were found to be significant, but negligible). It might be useful then to have a closer look at this section of the exam.

The Reading Comprehension section of the test comprises various questions which can generally be divided into three parts. The first part, which is the main part of the section, contains 28 questions and consists of six passages, ranging from 87 to 233 words, each followed by four to six multiple-choice questions. The second part which includes only four questions is called the Restatement part in which the test takers were supposed to read sentences followed by four choices and choose the best restatement of the sentences. The last part is called Coherence and has just three questions. For each item in this part, the test takers were required to read a paragraph in which one sentence had been removed and choose the choice that best completed the paragraph and made it coherent.

A cursory examination of the items with significant DIF in the Reading Comprehension section revealed that passages 1, 4, 5 and the Coherence part were mostly in favor of males. Though the DIF values were found to be "negligible", the finding that all of the significant DIF values related to the items following these three passages and the Coherent part favored males should not be taken as accidental. In general, the themes of the three passages favoring males can be summarized as "educational development", "education of children with disabilities", and "the International Bank", respectively. Nonetheless, it should be admitted that drawing any conclusions at this point is neither helpful nor reliable. In fact, a more careful and detailed analysis of the reading passages is needed before getting to any generalizations.

By and large, this study posed three research questions to investigate: First the presence, second the magnitude, and third the direction of any gender DIF in the items of the UTEPT. With respect to the first and second research questions, it was found that of the 100 items in the test, 39 items had a significant *p*-value for their test of significance; however, the detected DIF indices are considered "negligible" using the corresponding classification schemes for DIF effect size proposed by Zumbo and Thomas (1996) and Jodoin and Gierl (2001).

The direction of DIF can be estimated from the regression coefficients, or particularly from the gender group main effect *b*-weight or the slope parameter of the regression. In fact, the direction of DIF depends on the sign (positive or negative) of the group *b*-weight and how group membership is coded (B. Zumbo, personal communication, August 1, 2008). Although some efforts were made in the present study to explain the small differences between the performances of the two genders on some items (particularly on the Reading Comprehension section), it should be noted that the direction of DIF is interpretable only when there *is* a DIF at all. Therefore, since no DIF index at the item level was detected in the test, this question needs to be examined in a more comprehensive study which takes differential test functioning (see Pae & Park, 2006) or differential bundle (testlet) functioning into consideration (see Wainer et al. 1991).

The results of the study can bear more significance by taking one point into account. LR is a parametric DIF detection approach which is a response to the previous DIF techniques which could only screen uniform DIF such as standardization, MH or SIBTEST. In a study comparing LR and MH for detecting DIF, Rogers and Swaminathan (1993) concluded that the LR procedure was as powerful as the MH procedure in detecting uniform DIF, and more powerful than the MH in detecting non-uniform DIF (see also Gierl, Jodoin & Ackerman, 2000). In addition, as Duncan (2006) stated,

If LR DIF can detect non-uniform DIF better than the MH DIF method, and is as powerful at detecting uniform DIF as the MH DIF method, then the inclusion of an effect size would make LR DIF a very attractive choice as a DIF detection method. (p. 38)

Jodoin and Gierl (2001) also emphasized that while LR has comparable power to MH and SIBTEST in detecting uniform DIF, it is superior in power for detecting non-uniform DIF. In addition, Gierl, Rogers and Klinger (1999) found that "effect size measures [for MH, SIBTEST and LR] were highly correlated across DIF procedures except the measure for non-uniform DIF" (p. 15), which could only be assessed by LR. Altogether, these findings can provide tacit confirmation as to the superiority of LR over standardization, MH and SIBTEST.

There is another point which may add to the value of the findings of the study. In an attempt to examine the reliability of different DIF detection methods, Huang (1998) made a comparison between standardization, MH and LR methods and found that more items could be identified as exhibiting DIF by LR than the MH and standardization methods. That is, "items which were labeled as 'exhibiting DIF' by the MH and STD [i.e., standardization] methods could be identified as either uniform DIF or non-uniform DIF in the LR method" (p. 8). In other words, all the items labeled as "exhibiting DIF" by both the MH and standardization methods, were also detected to exhibit DIF by the LR method. It can be concluded, therefore, that in detecting DIF items "the LR method is more sensitive than the MH and STD methods" (p. 8), and that in comparison with MH and standardization, the LR

method tends to label the most items as exhibiting DIF. This is consistent with the findings of Gierl, Khaliq, et al. (1999), who found that in comparison with MH, which they referred to as "the conservative procedure" (p. 12), LR could flag a large number of items as exhibiting DIF. Also, as Gierl, Jodoin, et al. (2000) have found, LR has excellent Type I error rates which is a reassuring point for the researchers who choose LR as their DIF detection method.

On the whole, what can be inferred from this comparative discussion is that, generally, LR is more likely to flag an item with moderate or large DIF than the other two DIF detection methods which have been generally used at ETS viz standardization and MH. In other words, by utilizing LR, the researchers can be sure that they obtain a list of DIF items which might not be flagged as displaying DIF by either the MH or standardization procedures. Metaphorically, LR feels free to accuse an item of displaying DIF (see Duncan, 2006; Gierl, Jodoin, et al., 2000; Gierl, Rogers, et al., 1999; Jodoin & Gierl, 2001 and Rogers & Swaminathan 1993).

This, implicitly, can be considered as a reassuring point for the developers of the UTEPT. In other words, since the procedure which was used in the study is considered to be the strictest in comparison to standardization and MH, and since LR could not flag any item as displaying DIF in the test, the developers of the UTEPT can be convinced that it is most probable that the items of the test would not be flagged as displaying DIF if another study were to be conducted using standardization or MH. In this regard, a better plan for future studies may be to conduct another DIF analysis on the test utilizing item response procedures which seem to be the most efficient methods of all. For one thing, unlike MH and standardization, IRT is capable of screening non-uniform DIF which is important if the follow-up studies are to be comparative (see Kang & Cohen, 2007 for IRT model selection methods for dichotomous items). Generally, an ideal DIF study is the one which uses a combination of DIF techniques. Also, further research is needed to examine the differential functioning not of items but of the whole test or items bundles particularly the Reading Comprehension section. In addition, other studies can examine the existence of DIF for examinees with different academic

backgrounds or disciplines.

References

Breland, H., Lee, Y., Najarian, M. and Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups* (ETS-RR-04-05, Report 76). Princeton, NJ: Educational Testing Service.

Broer, M., Lee, Y., Rizavi, S., & Powers, D. (2005). *Ensuring the fairness of GRE writing prompts: Assessing differential difficulty* (GRE Board Report No. 02-07R, ETS RR-05-11). Princeton, NJ: Educational Testing Service.

Buck, G., Kostin, I. & Morgan, R. (2002). *Examining the relationship of content to gender-based performance differences in Advanced Placement Exams* (College Board Research Report No. 02-12, ETS Research Report 02-25). NY: College Entrance Examination Board.

Conoley, C. A. (2003). *Differential item functioning in the Peabody Picture Vocabulary Test – Third Edition: Partial correlation versus Expert judgment*. Unpublished doctoral dissertation, Texas A&M University, TX.

Curley, E. W. & Schmitt, A. P. (1993). *Revising SAT-Verbal items to eliminate differential item functioning* (College Board Research Report No. 93-2, ETS Research Report 93-61). NY: College Entrance Examination Board.

Dorans, N. & Kulick, E. (1986). "Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test". *Journal of Educational Measurement, 23*, 355 – 368.

Duncan, S. C. (2006). *Improving the prediction of differential item functioning: A comparison of the use of an effect size for logistic regression DIF and Mantel-Haenszel DIF methods*. Unpublished doctoral dissertation, Texas A&M University, TX.

Educational testing service fairness review guidelines (2003). Princeton, NJ: Educational Testing Service.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000, April). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and logistic regression when the proportion of DIF items is large.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA.

Gierl, M. J., Khaliq, S. N. & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. Paper presented at the Symposium entitled "Improving Large-Scale Assessment in Education" at the Annual Meeting of the Canadian Society for the Study of Education, Sherbrooke, Québec, Canada.

Gierl, M., J., Rogers, W. T., Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Montréal, Quebec, Canada.

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

Huang, C. (1998, April). Factors influencing the reliability of DIF detection methods. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Jodoin, M. G., & Gierl, M. J. (2001). "Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection". *Applied Measurement in Education, 14*, 329 – 349.

Kang, T. & Cohen A. (2007). "IRT Model Selection Methods for Dichotomous Items". *Applied Psychological Measurement, 31* (4), 331 – 358.

Kanjee, A. (2007). "Using logistic regression to detect bias when multiple groups are tested". *South African Journal of Psychology, 37*, 47 – 61.

Lawrence, I. & Curley, W. E. (1989). *Differential item functioning for males and females on SAT-Verbal Reading subscore items: Follow-up study* (ETS-RR-89-22). Princeton, NJ: Educational Testing Service.

Lin, J. & Wu, F. (2003). *Differential performance by gender in foreign language testing*. Poster session presented at the 2003 annual meeting of NCME, Chicago.

McNamara, T. & Roever, C. (2006). *Language Testing: The Social Dimension*. Massachusetts: Blackwell Publishing.

Pae, T. & Park, G. (2006). "Examining the Relationship between Differential Item Functioning and Differential Test Functioning". *Language Testing, 23* (4), 475 – 496.

Penfield, R. D. & Camilli, G. (2007). "Differential item functioning and item bias". In C.R. Rao & S. Sinharay (Vol. Eds.), *Handbook of statistics: Vol. 26* (pp. 125 – 167). Elsevier.

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.

Ryan, K. & Bachman, L.F. (1992). "Differential item functioning on two tests of EFL Proficiency". *Language Testing, 9*, 12 – 29.

Salkind, N. J. (Ed.). (2007). *Encyclopedia of measurement and statistics* (Vols. 1 – 3). Thousand Oaks, CA: Sage Publications.

Schmitt, A. P., Holland, P. W., Dorans, A. J. (1992). "Evaluating hypothesis about differential item functioning". In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281 – 315). Hillsdale, NJ: Lawrence Erlbaum Associates

Shealy, R., & Stout, W. (1993). "A Model-based Standardization Approach that Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DTF as Well as Item bias/DIF". *Psychometrika, 58*, 159 – 194.

Sheppard, R., Han, K., Colarelli, S. M., Dai, G. & King, D. W. (2006). "Differential Item Functioning by Sex and Race in the Hogan Personality Inventory". *Assessment, 13* (4), 442 – 453.

Sireci, S. G., Allalouf, A. (2003). "Appraising item equivalence across multiple languages and cultures". *Language Testing, 20* (2), 148 – 166.

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

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

Wainer, H., Sireci, S. G., Thissen, D. (1991). "Differential testlet functioning: Definitions and detection". *Journal of Educational Measurement, 28*, 197 – 219.

Woods, R. (1991). *Assessment and Testing: A Survey of Research*. Cambridge, MA: CUP.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. & Thomas, D. R. (1996). *A measure of effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.