Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. ٤, No. ٢, October ٢٠١٤

ISSN ٢٢٥١-٧٣٢٤

Detecting Gender DIF with an English Proficiency Test in EFL Context

Seyed Mohammad Reza Amirian[١], Seyed Mohammad Alavi[٢], Angel M. Fidalgo[٣]

Abstract

The aim of the present study is twofold. First, the paper investigated whether University of Tehran English Proficiency Test (UTEPT) manifested substantial gender Differential Item Functioning (DIF). Second, the flagged DIF items were subjected to a content analysis to determine underlying sources of DIF. Mantel-Haenszel (MH) and Logistic Regression (LR) as two popular methods of DIF detection were employed to analyze the data obtained from ١٥٥٠ test takers in ٢٠١٠. The findings indicated that even though ٢٨٪ of items were initially detected by MH and LR as displaying gender DIF, the effect size of DIF was mostly negligible. Moreover, the content analysis phase of the study showed that sometimes it is difficult to hypothesize the linguistic element causing DIF in items. However, humanities-oriented subjects were rated as favoring females and science-oriented subjects were rated as favoring males. Finally, a correlation index of .٩٠ manifested that MH and LR produce highly consistent DIF results. These findings are discussed and implications for test developers and DIF researchers are provided.

Key terms: *Fairness, DIF, Uniform DIF, Non-Uniform DIF, MH, LR, UTEPT*

١. Introduction

The present Differential Item Functioning (DIF) study is an attempt to examine University of Tehran English Proficiency Test (UTEPT) for gender DIF in order to ensure test fairness and detect the potential sources of bias. Test fairness is an issue of utmost importance in language testing which is closely related to test validity and test validation (Kunnan, ٢٠١٠; Xi, ٢٠١٠). DIF is said to be present when examinees from different groups show differing probabilities of success on an item after they are matched on the underlying ability that the item is intended to measure (Zumbo, ١٩٩٩).

It is to be emphasized that finding DIF in an item does not necessarily imply that the item is biased, that is, unfair to one of the groups (Angoff, ١٩٩٣). DIF is a necessary but not sufficient condition for bias. An item may show DIF but not be biased if the difference is because of actual differences in the groups' ability needed to answer the item, for example, if one group is high proficiency and the other low proficiency: The low proficiency group would necessarily score much lower. Only where the difference is caused by construct-irrelevant factors can DIF be viewed as bias. In other words, an item that shows DIF needs to be investigated further to uncover the reasons for its differential functioning.

[١] *Hakim Sabzevari University, Sabzevar, Iran. Email: Sm.Amirian@hsu.ac.ir*

[٢] *University of Tehran, Tehran, Iran. Email: Mohammed.Alavi@gmail.com*

[٣] *Universidad de Oviedo, Plaza de Feijoo, s/n, Oviedo, ٣٣٠٠٣, Spain. Email: Fidalgo@umiovi.es*

ISSN ٢٢٥١-٧٣٢٤

As one of the earliest and most popular DIF detection techniques, Mantel-Haenszel (MH) has been widely used in DIF studies (Dorans & Kulick, ٢٠٠٦; Hambleton & Rogers, ١٩٨٩; Roever, ٢٠٠٧). Interest in using the MH procedure was shown after publication of the paper by Holland and Thayer (١٩٨٨). The primary reasons for the popularity of MH are attributed to its computational simplicity, ease of implementation, associated test of significance and capacity for detecting DIF with small sample sizes (Fidalgo, Alavi, & Amirian, ٢٠١٤; Fidalgo, Ferreres, & Muniz, ٢٠٠٤; Fidalgo, Hashimoto, Bartram, & Muniz, ٢٠٠٧; Rogers & Swaminathan, ١٩٩٣).

As DIF detection is method dependent, DIF researchers highly recommended to use more than one method in order to cross-validate DIF results (e.g., Ferne & Rupp, ٢٠٠٧).Logistic Regression (LR) is employed in this study as a second method to triangulate DIF detection methods. LR has recently been widely used for DIF detection in many disciplines including language testing (Alavi, Amirian & Rezaee, ٢٠١١; Fidalgo, Alavi, & Amirian, ٢٠١٤; Breland, Lee, Najarian, & Muraki, ٢٠٠٤; Kim, ٢٠٠١). LR analysis has been used mainly to study group effect for dichotomously scored test items (Swaminathan& Rogers, ١٩٩٠), but French and Miller (١٩٩٦) demonstrated that this procedure can be extended for polytomous items as well. Some other reasons for the popularity of LR are that it allows modeling of Uniform DIF (UNDIF) and Non-Uniform (NUDIF) and requires less complicated computing than IRT-based analyses.

Although gender DIF has been extensively researched in first language contexts (Aryadoust, Goh, &Kim, ٢٠١١; Li & Suen, ٢٠١٣; Ryan & Bachman, ١٩٩٢; Zhang, Dorans &Matthews-Lopez, ٢٠٠٥) few studies have dealt with DIF in EFL/ESL settings (Alavi, Rezaee, & Amirian, ٢٠١١, Pae, ٢٠٠٤; Park & French, ٢٠١٣, Shimizu & Zumbo, ٢٠٠٥). In fact, the literature on DIF studies on tests that are developed and administered outside the United States is critically meager (Ferne & Rupp, ٢٠٠٧). To bridge this gap, and in an attempt to validate UTEPT, the present DIF study uses MH and LR to find out whether UTEPT as a test developed and administered in Iran shows substantial DIF in favor of a specific gender group. Moreover, by conducting a content analysis of DIF items, the possible underlying linguistic sources of DIF are examined. Therefore, the study addresses the following questions:

Q١. Do items of UTEPT show substantial DIF in favor of males or females after matching on ability?
Q٢. To what extent are MH and LR gender DIF findings comparable?
Q٣. Does content analysis of DIF items indicate bias in UTEPT?

٢. Previous Gender DIF Studies

Many studies have recently been conducted to detect DIF with language tests (e.g. Alavi, Rezaee, & Amirian, ٢٠١١, Geranpayeh & Kunnan, ٢٠٠٧,Pae& Park, ٢٠٠٦) and DIF investigation has become "a key component of validity studies in virtually all large-scale assessments" (Penfield & Camilli, ٢٠٠٧, p.١٢٥). Nevertheless, research into gender DIF has indicated conflicting results (Mielikainen, ١٩٨٨; Tannen, ١٩٩٠). Some of these studies are reviewed here.

In an early DIF study, Ryan and Bachman (١٩٩٢) investigated differential performance on the TOEFL and the FCE. They found little evidence that males and females performed differently at the item level on either test. Wainer and Lukhele (١٩٩٧) also reported that the reading comprehension testlets of TOEFL showed essentially no differential functioning by

ISSN ٢٢٥١-٧٣٢٤

gender. This is in contrast with the findings ofCarlton and Harris (١٩٩٢) who in their analysis of gender DIF using MH and LR found that overall the female group performed differentially better than a matched male group.

Many studies have also focused on DIF investigation at skills level. In their investigation of the comparability of TOEFL writing prompts, Breland et al. (٢٠٠٤) employed LR procedures to estimate prompt difficulty and gender effects. They found that open-ended questions generally favored female test takers, which they hypothesized might have been due to gender differences in reasoning and cognitive processes.

MELAB listening test was examined for gender DIF by Aryadoust, Goh, and Kim (٢٠١١) using Rasch measurement. The result of a *t*-test UDIF analysis showed that two test items displayed substantive DIF favoring different gender subgroups; and NUDIF analysis revealed several test items with significant DIF, many of which favored low-ability male test takers.

Among very few DIF studies with a non-U.S sample, Gafni (١٩٩١) used the Standardization and the MH procedures to examine gender DIF. It was found that only three out of a total of ٥٠ English items were identified as showing gender DIF. Similarly, Lin and Wu (٢٠٠٣) in their study of EPT in China found that out of a total of ١٢٠ items, only two items revealed C-level or large DIF and in fact ٨٩٪ of the test items manifested "no" DIF whatsoever.

To sum up, it can be said that research into gender DIF has produced inconsistent results and the issue needs further investigation using more elaborate methodologies. To this end, as a DIF study in an EFL context, the current study focuses on DIF analysis of UTEPT. In a previous DIF investigation with UTEPT focusing on academic discipline differences, Alavi, Rezaee, and Amirian (٢٠١١) found that overall UTEPT shows no DIF for humanities vs. science and engineering groups. Two previous studies (Rezaee & Shabani, ٢٠١٠, and Karami, ٢٠١١) also examined earlier versions of UTEPT for gender DIF. In the first study, Rezaee and Shabani (٢٠١٠) studied ٦,٥٥٥ examinees and found that ٣٩ of the ١٠٠ items in the test displayed significant gender differences. However, these group differences were viewed as "negligible". This finding was confirmed by Karami (٢٠١١) who utilized Rasch model to investigate gender DIF in UTEPT. The results of his study indicated that ١٩ items were functioning differentially for the two groups. Only ٣ items, however, displayed DIF with practical significance. These two studies, however, did not investigate the content of DIF items to uncover underlying causes of DIF. To fill this gap, the present study investigates DIF items in all grammar, vocabulary and reading subtests of UTEPT both quantitatively and qualitatively in order to look into the linguistic sources of DIF.

٢.١. Linguistic Causes of DIF

There is a scarcity of research into content-based reasons of DIF in DIF literature. In one of the early DIF studies investigating causes of DIF, Kim (٢٠٠١), examined DIF across a sample of ١٠٣٨ Asian and European participants in a speaking test using Likelihood ratio test and the LR methods. The results showed that 'grammar' and 'pronunciation' functioned differentially across the two groups. A content analysis of the DIF items suggested that the types and the numbers of scoring scales might influence the test validity.

In another study, Uiterwijk and Vallen (٢٠٠٥) investigated linguistic sources of item bias for second generation immigrants in Dutch tests employing IRT and MH. An investigation of sources of DIF was carried out which was followed by an in-depth analysis

ISSN ٢٢٥١-٧٣٢٤

of DIF items by the researchers and by external experts. The findings indicated that it was not easy for judges to decide if the DIF items were biased or not, but the description of the domains the items claim to measure turned out to be very informative and helpful in getting a final decision. While ١٧.٤٪ of the items showed DIF, the researchers eventually concluded that only ٤٪ of all the items were biased.

In a similar study, Geranpayeh and Kunnan (٢٠٠٧) concentrated on the classification of linguistic sources of DIF of immigrant students. They investigated whether the test items on the listening section of the Certificate in Advanced English examination functioned differently for test takers from three different age groups. The main results showed that although a few items were detected based on statistical and content analyses procedures, expert judges could not clearly identify the sources of differential item functioning for the items. As it is essential to identify source of DIF, we employ a DIF questionnaire in the content analysis phase of the study to elicit expert panels' judgments (Ferne & Rupp, ٢٠٠٧) to uncover underlying sources of linguistic bias in UTEPT.

٣. Methodology

The data for the present study were gathered from ١٥٥٠ test takers who took UTEPT in ٢٠١٠. The sample was divided into a reference group of ٨٩٩ male and a focal group of ٦٥١ female test takers with various age ranges. The participants were all PhD candidates seeking to sit for PhD exams of the University of Tehran. After generating the data, the descriptive statistics was computed using SPSS. The results indicated a slight difference between the mean score of male (*M*= ٥٠.٩٦, *SD*= ١٢.٧٠) and female (*M*=٥٠.٦١, *SD*= ١٣.١٤) examinees. The effect size of mean difference is estimated small according to Cohen's (١٩٨٨) test (*d*= .٠٢٧) .The reliability of the data was also estimated ٠.٨٨ using Cronbach's alpha which indicates that the test enjoys high reliability.

MH statistic can be calculated using easily accessible general statistical softwares (SPSS, SAS) or more specific packages (MHDIF: Fidalgo, ١٩٩٤; DIFAS: Penfield, ٢٠٠٥, ٢٠٠٩). In the present study, DIFAS ٥.٠ (Penfield, ٢٠٠٩) was used for MH DIF detection. In addition to reporting MH □ (Holland and Thayer, ١٩٨٨), DIFAS has the advantage of reporting a test of effect size based on ETS classification scheme (Zieky, ١٩٩٣).

In order to run LR as a second method of DIF analysis, the Nagelkerke'sSPSS syntax for nominal data (Zumbo, ١٩٩٩) was utilized. The LR procedure uses the item response (٠ for incorrect response or ١ for correct response) as the dependent variable, with grouping variable (dummy coded as ١=male, ٢=female), total score (characterized as variable TOT) and a group by TOT interaction as independent variables. This appears in the following equation: $Y = b_0 + b_1 TOT (\theta) + b_2 Group (g) + b_3 TOT * Group (\theta*g)$.

After computing the two-degree-of-freedom □□valueas a test of DIF significance, Zumbo-Thomas'(١٩٩٧) effect size measure was calculated. Jodoin and Gierl's (٢٠٠١) more conservative criteria were also employed as a second test of DIF magnitude. Items which showed DIF were subjected to further content analysis using a DIF questionnaire developed by Geranpayeh and Kunnan (٢٠٠٧). Accordingly, two content experts were asked to rate the suitability of the test items for each gender group using a questionnaire with a ٥-point scale. It was expected that if evidence exists that the items detected with DIF were advantaging a particular group of test takers, then the items may be biased.

٤. Results

ISSN ٢٢٥١-٧٣٢٤

٤. ١. *Mantel-Haenszel Results*

The aim of this study was to assess the possibility of existence of DIF in UTEPT, to check the comparability of MH and LR DIF findings, and to examine the content of DIF items for potential sources of linguistic bias.

The result of MH analysis for males and females is presented in Table ٢. For the sake of convenience, only items that are flagged as DIF items are shown in Table ٣.Out of a total of ١٠٠ UTEPT items, ٣١ items (٣١٪) are flagged with MH gender DIF. According to ETS classification scheme, however, ٢٤ items displayed category A (negligible) effect size and only ٧ items manifested category B (Moderate) size DIF while no item was categorized as category C (large) DIF magnitude.

Table ٢.MH Gender DIF Results

Item	Section	Favored	MH CHI	MH LOR	ETS
١	G	F	١٠.٧٠٩	-٠.٤٣٧	B
٢	G	F	٦.٠٥٣	-٠.٢٨٣	A
٣	G	F	١٠.٦٨٢	-٠.٣٦٨	A
٨	G	F	٦.١٢٣	-٠.٢٧٦	A
٩	G	F	٩.٨٦٦	-٠.٣٨٧	A
١٠	G	F	٧.٦٠٣	-٠.٣٧١	A
١٣	G	F	٧.٩٠٨	-٠.٣٢٧	A
٢٠	G	F	١٦.٧٢٣	-٠.٤٧٣	B
٢٢	G	F	٢٥.٨٨٧	-٠.٦٢٥	B
٢٤	G	F	٧.٣٩٩	-٠.٣٤٢	A
٢٥	G	F	٧.٩٤٢	-٠.٣٥٩	A
٢٦	G	F	٨.٦٥٤	-٠.٥٢٢	B
٢٨	G	F	٦.٦٨٠	-٠.٣٣٧	A
٢٩	G	F	٠.١٤٠	-٠.٠٤٧	A
٣٥	V	M	٤.٨٥٣	٠.٢٤٨	A
٣٨	V	M	١٢.٥٢١	٠.٤٩٥	B
٤١	V	M	٩.٩٥٩	٠.٤٦٤	B
٤٤	V	M	٧.١٧٠	٠.٣١٠	A
٤٨	V	M	٢.٥٩٨	٠.٢٢١	A
٤٩	V	M	٥.٤٥٨	٠.٢٩٤	A
٥٠	V	M	٩.٩٤٨	٠.٤٠٠	A
٥١	V	M	٠.٤٨٥	٠.٠٨٢	A
٥٢	V	M	١٦.٦٠	٠.٥٠٥	B
٥٤	V	M	٦.٧٥٣	٠.٣٠٩	A
٥٩	V	M	٨.١٧١	٠.٣٢٦	A
٨١	R	F	٥.٨٠٧	-٠.٣٧٠	A
٩٠	R	F	٠.٣٩٢	-٠.٠٨	A

ISSN ٢٢٥١-٧٣٢٤

٩١	R	M	٦.٦٧١	٠.٢٩٣	A
٩٢	R	F	٢.٧١١	-٠.٢٠٣	A
٩٥	R	M	٥.٢٨٠	٠.٣٢٤	A
٩٧	R	M	٠.٤٠٢	٠.٠٧٦	A

Notes. *$p < .٠٥$; G= Grammar; V= Vocabulary; R= Reading; F = Female; M = Male; A = negligible DIF; B = moderate DIF

MH CHI = Mantel-Haenszel Chi-Square statistic which is distributed as chi-square with one degree of freedom.

MH LOR= Mantel-Haenszel Common Log-Odds Ratio. Positive values indicate DIF in favor of males and negative values indicate DIF in favor of females.

DIF items come from different sections of UTEPT. Out of ٣١ MH DIF items, ١٤ items belong to the grammar section, ١١ items to the vocabulary section and ٦ items to the reading comprehension section. This shows that the grammar section displays more gender DIF items than vocabulary and reading sections. In fact, both vocabulary and reading sections contain ١٧ DIF items while grammar section alone incorporates ١٤ DIF items.

In terms of the direction of DIF, ١٧ items favored females and ١٤ items favored males. It was also indicated that the grammar section primarily worked to the advantage of females, vocabulary section to the advantage of males and reading comprehension section neither to the advantage of males nor females.

٤.٢. Logistic Regression Results

The findings of LR DIF are summarized in Table ٣. It should be noted that only items with significant DIF values at ٠.٠٥ level of significance are included in the table. It appears that the two-degree-of-freedom chi-squared (χ^2) test of the significance for DIF is significant for ٢٩ items. Of these ٢٩ items, ١٤ items are detected in the grammar section, nine items in the vocabulary section and only six items in the reading comprehension section.

The obtained R٢ values reveal that gender DIF is predominantly of uniform nature on UTEPT. Of a total of ٢٩ DIF items, ٢٥ items displayed UDIF and only four items displayed NUDIF.١٩ of UDIF items favored the female group and only ٦ items favored the male group. Additionally, based on Copella and Sireci's (٢٠٠٩) guidelines, three NUDIF items (٥١, ٩٠, and ٩٢) were found to favor females and only one item (٤٨) appeared to favor males. All in all, out of ٢٩ DIF items, ١٧ items favored females and ١٢ items favored males. Concerning DIF effect size, based on Jodoin and Gierl's (٢٠٠١) conservative classification scheme, it is observed that all obtained R٢ values are category A that is smaller than .٠٣٥. This means that all LR DIF items manifest a negligible DIF magnitude based on guidelines proposed by both Zubmo and Thomas (١٩٩٧) and Jodoin and Gierl (٢٠٠١).

ISSN ٢٢٥١-٧٣٢٤

Table ٣. LR Gender DIF Results

Item	Section	Favored	UDIF	NUDIF	DIF	size	χ^2	P*	Category
١	G	F	.٠١٣	.٠٠٢	.٠١٥		١٧.٧٤٠	.٠٠٠	A
٢	G	F	.٠٠٧	.٠٠٠	.٠٠٧		٩.٢٤٢	.٠١٠	A
٣	G	F	.٠٠٨	.٠٠٠	.٠٠٨		١٠.١٦٤	.٠٠٦	A
٨	G	F	.٠٠٤	. ٠٠٢	.٠٠٦		٧.١٥٢	.٠٢٨	A
٩	G	F	.٠٠٥	.٠٠١	.٠٠٦		٨.٩٢٥	.٠١٢	A
١٠	G	F	.٠١٠	.٠٠٠	.٠١٠		٩.٥٦١	.٠٠٨	A
١٣	G	F	.٠٠٧	.٠٠٠	.٠٠٧		٧.٦٩٨	.٠٢١	A
٢٠	G	F	.٠١٥	.٠٠١	.٠١٥		٢٣.٤٧٦	.٠٠٠	A
٢٢	G	F	.٠١٩	.٠٠٢	.٠٢١		٢٨.٢٨٠	.٠٠٠	A
٢٤	G	F	.٠٠٦	.٠٠٠	.٠٠٦		٧.٨٩٣	.٠١٩	A
٢٥	G	F	.٠٠٥	.٠٠١	.٠٠٦		٦.٧٩٠	.٠٣٤	A
٢٦	G	F	.٠٠٩	.٠٠٠	.٠٠٩		١٠.٠٧٦	.٠٠٦	A
٢٨	G	F	.٠٠٧	.٠٠١	.٠٠٨		١٠.٢٨١	.٠٠٦	A
٣٥	V	M	.٠٠٤	.٠٠٣	.٠٠٧		٨.٥٠٩	.٠١٤	A
٣٨	V	M	.٠١٢	.٠٠١	.٠١٣		١٣.٥٠٤	.٠٠١	A
٤٤	V	M	.٠٠٥	.٠٠٠	.٠٠٥		٦.١٣٩	.٠٤٦	A
٤٨	V	M	.٠٠٣	.٠٠٧	.٠١٠		٩.٦٧٤	.٠٠٨	A
٤٩	V	M	.٠٠٤	.٠٠٢	.٠٠٦		٨.٢٢٦	.٠١٦	A
٥٠	V	M	.٠٠٦	.٠٠٠	.٠٠٦		٧.٩٦٥	.٠١٩	A
٥١	V	F	.٠٠٠	.٠٠٥	.٠٠٥		٦.٣٢٢	.٠٤٢	A
٥٢	V	M	.٠١١	.٠٠١	.٠١٢		١٥.٨٦١	.٠٠٠	A
٥٤	V	M	.٠٠٥	.٠٠٣	.٠٠٦		٧.٠٠٥	.٠٣٠	A
٥٩	V	M	.٠٠٦	.٠٠١	.٠٠٩		١١.٨٤١	.٠٠٣	A
٦٧	R	M	.٠٠٦	.٠٠٠	.٠٠٧		٧.٩٦١	.٠١٩	A
٨١	R	F	.٠٠٥	.٠٠٥	.٠٠٥		٦.٠١١	.٠٥٠	A
٩٠	R	F	.٠٠١	.٠٠٠	.٠٠٦		٦.٦٥١	.٠٣٦	A
٩١	R	M	.٠٠٦	.٠٠٦	.٠٠٦		٨.٠٦٠	.٠١٨	A
٩٢	R	F	.٠٠١	.٠٠٤	.٠٠٧		١٠.١٢١	.٠٠٦	A
٩٥	R	M	.٠٠٤	.٠٠١	.٠٠٨		٨.٨٢٧	.٠١٢	A

Notes. *p <.٠٥; G= Grammar; V= Vocabulary; R= Reading; F = Female; M = Male; A = negligible DIF

ISSN ٢٢٥١-٧٣٢٤

As far as the comparability of MH and LR DIF findings is concerned, it was found that MH detected two more gender DIF items (٣١) than LR (٢٩) (Table ٤). This finding contrasts with the dominant view in DIF literature that LR detects more DIF items in comparison to MH due to its capability of detecting both UDIF and NUDIF (Hidalgo, & López-Pina, ٢٠٠٤; Rogers & Swaminathan, ١٩٩٣). Nevertheless, Phi correlation for nominal data was run to compare the performance of MH and LR methods in flagging similar DIF items in all ١٠٠ items. The flagged items by either method were dummy coded as ١ and the unflagged ones were dummy coded as ٢. Phi correlation value of .٩٠ indicated a highly significant correlation between the findings of the two methods.

Table ٤.Comparison of Gender DIF Results in Sections of UTEPT

	Number of items detected by MH only	Number of items detected by LR only	Number of items detected by both methods
Grammar	١٤	١٣	١٣
Vocabulary	١١	١٠	١٠
Reading Comprehension	٦	٦	٥
Total	٣١	٢٩	٢٨

In terms of specific sections of DIF items, MH detected ١٥ gender DIF items in grammar, ١٠ items in vocabulary, and six items in reading comprehension sections while LR flagged ١٤ DIF items in grammar, nine items in vocabulary and six items in the reading comprehension sections. This shows a great consistency between MH and LR findings even in their DIF detection in sections of UTEPT. As far as the magnitude of DIF is concerned, however, it was found that all DIF items detected by LR method displayed a negligible or type-A effect size while MH detected ٧ items with moderate or type-B effect size.

٤.٣. Content Analysis of DIF Items

٢٨ DIF items detected by both LR MH and LR methods were examined based on elicited responses from two expert judges. The results of content analysis for gender DIF are reported below.

Grammar Items. It is interesting that grammar section of UTEPT contains greater number of DIF items than vocabulary and reading comprehension sections. Among ١٤ items flagged with DIF in the grammar section, seven items (١, ٢, ٣, ٩, ١٠, and ١٣)came from the *fill in the blank structure subsection* four items (٢٠, ٢٢, ٢٤, and ٢٥) came from the *written expression subsection* and three items (٢٦, ٢٨, and ٣٥)came from the *grammar in context* (*cloze test*) subsection of the test. For many DIF items in this section, neither of the content specialists could hypothesize the sources of DIF. Item ١ represents such items.

ISSN ٢٢٥١-٧٣٢٤

***Item* ١.** *Having been …….. the prize, the professor continued working hard on his project.*

A. awarded *B. award*
C. awarding *D. the award*

As it is shown in the Item Characteristic Curve (ICC) of this item (Figure ١), line ١ and ٢ do not cross which means that this item shows UDIF favoring females at all levels of proficiency**.** This item shows moderate DIF magnitude in favor of female test takers. Content experts believed there is no clue in such a short grammar item as to why this item is working to the advantage of females.

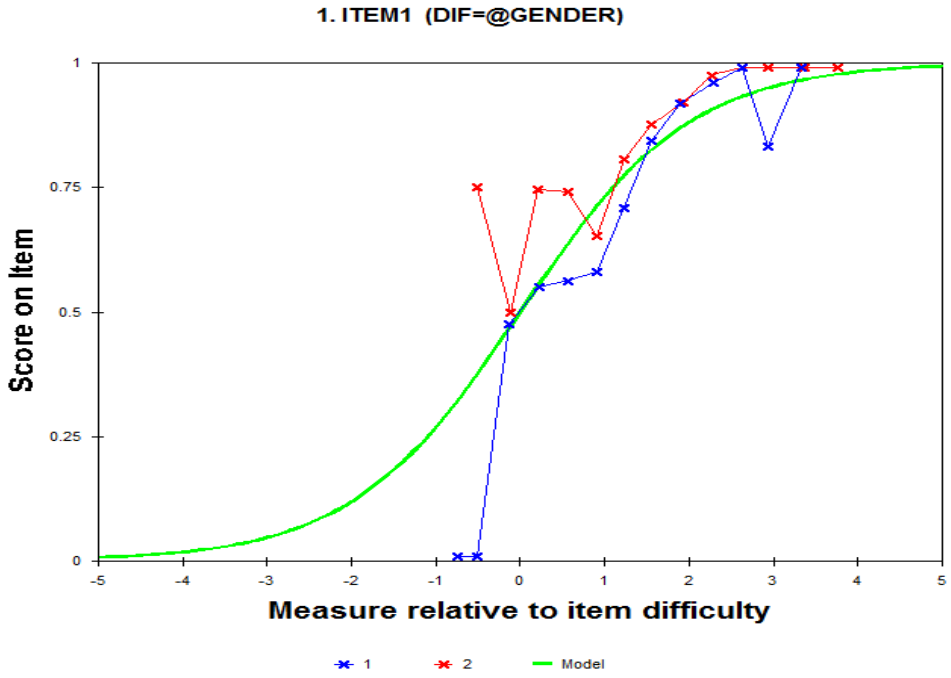

Figure ١.ICC for item ١
Note. ١= the performance of males; ٢ = the performance of females

Item ٢٦, which belongs to the *grammar in context* (*cloze test*)subsection, is an example of very few grammar items that content specialists unanimously rated as favoring females.

Item ٢٦. *Without regular supplies of some hormones, our capacity to behave would be seriously impaired; without others we would soon die. Tiny amount of some hormones can….. ٢٦……our moods and our actions.*

A. Modification *B. Modifying*
C. Modify *D. Modified*

This item was shown to advantage examinees in the female group by both MH and LR methods. The expert judges believed that since females show more interest in topics such as *human biology*, they are systematically favored by this item. One of the judges also pointed out that words such as "*behave and mood*" in the item indicate social interactions which is a topic of women's interest.

ISSN ٢٢٥١-٧٣٢٤

Vocabulary Items. Out of ٣٠ DIF items, ٩ items in the vocabulary section were flagged with DIF by both methods. The underlying cause of many of DIF items in vocabulary section was attributed to the texts in which these words were more likely to occur. For example, one of experts commented that in item ٣٨ the word "*fundamental*" (the correct choice) is more likely to be found in scientific texts which are of more interest to males.

Item ٣٨. *Analytic tools enable one to get at the most fundamental logic of any discipline.*
A. enforced *B. essential*
C. established *D. escorted*

Reading Comprehension Items. Out of ٣٥ items in this section, only six items (٦٪ of UTEPT items) were found to manifest DIF (items ٨١, ٩٠, ٩١, ٩٢, ٩٥ and ٩٧). Among these six DIF items, item ٨١ advantages females and comes from passage three which is on *history*. Content specialist commented that the content of this passage is responsible for differential performance because females score higher on humanities-oriented passages. Items ٩٠, ٩١, ٩٢ and ٩٥ (favoring females, males, females and males respectively) come from passage six on *painting art*. Item ٩٢ which is a "*scanning details*" question is presented below. The ICC for this item also appears in Figure ٢.

Item٩٢. *According to the passage, which of the following was one of the distinguishing characteristics of Impressionist painting?*
A. *The emphasis on people rather than nature scenes*
B. *The way the subjects were presented from multiple angles*
C. *The focus on small solid objects*
D. *The depiction of the effects of light and color*

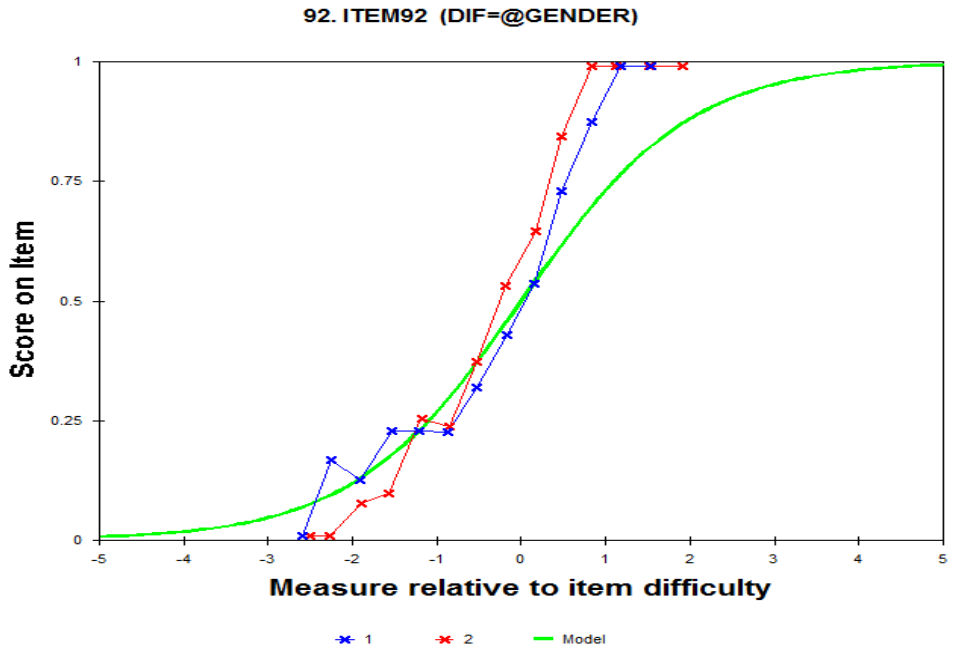

Figure ٢.ICC for item ٩٢

ISSN ٢٢٥١-٧٣٢٤

Note. ١= the performance of males; ٢= the performance of females

The ICC for this NUDIF item indicates that line ١ and ٢ cross at some points on the curve. It means that there is an interaction between proficiency level and grouping factor: the male group has a greater probability of answering an item correctly within the lower score range and the female group has a higher probability of answering an item correctly in the higher score range. However, overall the item favors females.

Content experts found *history* topic more humanities-oriented and as a result rated this item along with items ٩٠ (*main idea question*), ٩١ (*vocabulary question*), and ٩٥ (*scanning details question*) in favor of females. However, according to DIF findings items ٩١ and ٩٥ favored males.

٥. Discussion

The results of the study suggested that; altogether, MH flagged ٣١ DIF items and LR flagged ٢٩ items. Moreover, ٢٨ items were flagged with gender DIF by both methods in grammar (١٤ items), vocabulary (٩ items), and reading comprehension sections (٥items). This indicates that MH and LR produce consistent result which is confirmed by a high correlation index of .٩٠. This is in line with the findings of previous DIF studies that various DIF detection methods show a close correlation (e.g. Rogers & Swaminathan, ١٩٩٣).

The percentage of flagged DIF items in the present study (٢٨٪) is smaller than Rezaee and Shabani's (٢٠١٠) finding (٣٩٪) and larger than Karami's (٢٠١١) finding (١٩٪) with earlier editions of UTEPT. The highest number of DIF items were detected in the grammar section while reading comprehension section incorporates the lowest number of DIF items which is against our expectations as the reading section of many tests is found to be responsible for differential performance of them and this skill is the one that has been most extensively studied for DIF (e.g. Jiao & Chen, ٢٠١٤; Li & Suen, ٢٠١٣; Pae, ٢٠٠٤).

A closer look at the direction of DIF items reveals that surprisingly most of the items that favor females belong to the grammar section (١٤ items) and only three items belong to the reading section while no item in vocabulary section is in favor of females. In addition, most items that favor males come from the vocabulary section (١٠ items) while only one grammar item and three reading comprehension items favor males. Thus, it is concluded that overall the grammar section of UTEPT is in favor of females, the vocabulary section is in favor of males and the reading section is neither in favor of females nor males. This finding indicates that the developers or users of UTEPT should be aware of the test-takers for whom the test is intended. For example, high scores on the grammar subtest may not provide sufficient information to permit inferences about a female test taker's overall English ability since it showed DIF in advantage of females.

The findings of the content analysis phase of the study indicated that determining underlying causes of gender DIF was very hard for content specialists especially with short grammar and vocabulary items. Conoley (٢٠٠٣) also pointed out that it is not always clear which element in an item causes DIF. Nonetheless, for ٧٪ of the items it was the content of the item that was deemed responsible for differential performance. Experts considered humanities-oriented topics in favor of females and science-oriented topics in favor of males giving support

ISSN ٢٢٥١-٧٣٢٤

to Doolittle and Welch (١٩٨٩) who reported that females scored higher than males with humanities-oriented passages, but lower than males with science-oriented passages. This can be attributed to the educational system in Iran in which up to the recent years women showed more interest in studying in humanities majors and men showed greater interest in science and engineering majors. Of course, this pattern has changed lately and over ٦٠٪ of classes at universities are currently occupied by girls in all majors including science and engineering.

The findings of this study provide some implications for DIF researchers. The test of effect size in this study revealed that out of initially ٢٨ flagged items, MH only detected ٧ items with moderate magnitude while LR detected no sizable DIF magnitude whatsoever. This finding supports the results of Rezaee and Shabani's (٢٠١٠) who did not detect any sizable DIF item with UTEPT. It also highlights the significance of reporting the effect size in DIF studies(Cohen, ١٩٨٨; Kirk, ١٩٩٦). Although considering the direction of DIF can give us a lot of information about the behavior of DIF items, if the magnitude of DIF along with the direction of DIF is considered, a better picture of misbehaving items is obtained.

UTEPT developers and administrators may also benefit from the findings of this study. First, considering the large number of applicants who take this test each year, it is obvious that a DIF study needs to be a necessary part of the validation process of this test. Second, in order to promote gender equity, test developers should make sure the passage topics appeal to both genders and cover a wide range of academic subjects. Ultimately, although the magnitudes of DIF items are relatively small, the present pool of UTEPT items contains ٢٨ items that produce statistically significant gender differences. Although the observed differences are not large by accepted statistical standards, UTEPT developers should adopt a policy by specifying what levels of difference should result in items being revised or dropped from active administration.

٥.١. Conclusion and Suggestions for Further Research

This study made effort to make a contribution to the DIF literature by providing information about DIF with an Iranian sample. Since this is a test developed and administered in an EFL context, it may have some specific features that are observed only in Iranian contexts and is not common across nationalities. Even though initially about ٢٨٪ were consistently flagged as displaying DIF by MH and LR due to significant uniform or non-uniform group effects, their effect sizes were far too small for most of them to render the test unfair. That is, the item score differences between males and females compared in this study primarily seem to be because of item impact rather than group difference attributable to a construct irrelevant factor inherent in items or item bias. This fact reveals the importance of follow-up content analysis phase in DIF studies as not every DIF item is necessarily biased.

The present study was limited to the study of DIF at item level. Future research could address the differential functioning of the whole UTEPT test or sections of the test (bundles), particularly the grammar and vocabulary sections of UTEPT that were found to contain the largest number of DIF items in the current study.

Moreover, the current study analyzed the content of items only for linguistic sources of DIF. It is quite possible, however, that the DIF source is of a nonlinguistic nature. Therefore, other studies could be carried out looking for nonlinguistic sources of variance that contribute to differential performance of tests including item type, examinees' sociocultural background and

ISSN ٢٢٥١-٧٣٢٤

examinees' age. Finally, in the present study the researcher only consulted expert judges in order to identify the sources of DIF. Since identifying the potential sources of DIF in a test is a multidimensional process, interviews with examinees could be a promising line of inquiry into the causes of DIF.

Acknowledgments
Dr. Fidalgo's work in this article was supported by the Spanish Ministry of Economy and Competitiveness (grant number PSI٢٠٠٩-٠٨٥٢٩).

References

Alavi, S. M., Rezaee, A. & Amirian, S. M. R. (٢٠١١).Academic discipline DIF in an English language proficiency test. *Journal of English Language Teaching and Learning*, ٥(٧): ٣٩-٦٥

Angoff, W. H. (١٩٩٣). Perspectives on differential item functioning methodology. In P. W. Holland and H. Wainer (Eds). *Differential Item Functioning* (pp.٣-٢٥). New York: Routledge

Aryadoust, V., Goh, C. C. M. & Kim, L. O. (٢٠١١). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, ٨(٤): ٣٦١–٣٨٥.

Breland , H., Lee, Y., Najarian, M. & Muraki, E. (٢٠٠٤). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups* (ETS-RR-٠٤-٠٥, Report ٧٦). Princeton, NJ: Educational Testing Service.

Carlton, S. T., & Harris, A. M. (١٩٩٢).*Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons*. Princeton, NJ: Educational Testing Service.

Cohen. J. (١٩٨٨). Statistical *power analysis for behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Association.

Conoley, C. A. (٢٠٠٣). *Differential item functioning in the Peabody Picture Vocabulary Test – Third Edition: Partial correlation versus Expert judgment*. PhD Thesis, Texas A and M University, TX.

Copella, J., & Sireci, S. G. (٢٠٠٩). *Interpreting non-uniform DIF*. Poster for NCME ٢٠٠٩ Graduate Student Poster Session. University of Massachusetts Amherst.

Doolittle, A., Welch, C. (١٩٨٩). Gender differences in performance on a college level achievement test. Iowa City, IA.: American College Testing Program.

Dorans, N.J., Kulick, E. (٢٠٠٦). Differential item functioning on the Mini-Mental State Examination: An application of the Mantel-Haenszel and standardization procedures. *Medical Care,* ٤٤(١١, Suppl.٣), S١٠٧-S١١٤.

Ferne, T., Rupp, A. A. (٢٠٠٧). A synthesis of ١٥ years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, ٤(٢)*: ١–٣٦.

Fidalgo, A.M. (١٩٩٤). MHDIF: A computer program for detecting uniform and non-uniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological*

ISSN ٢٢٥١-٧٣٢٤

Measurement, ١٨, ٣٠٠-٣١٥.

Fidalgo, A. M., Alavi, S. M. & Amirian, S. M. R. (٢٠١٤). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing*, first published on April ١١, ٢٠١٤ as doi:١٠.١١٧٧/٠٢٦٥٥٣٢٢١٤٥٢٦٧٤٨ pp. ١-١٩

Fidalgo, A. M, Hashimoto, K., Bartram, D. & Muñiz J. (٢٠٠٧) Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *Journal of Experimental Education, ٧٥*, ٢٩٣-٣١٤.

Fidalgo, A. M., Ferreres, D.& Muniz, J. (٢٠٠٤). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for type I and type II rates. *The Journal of Experimental Education, ٧٣*(١):٢٣-٣٩

French, A.W., & Miller, T.R. (١٩٩٦). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, ٣٣*(٣): ٣١٥-٣٣٢.

Gafni , N. (١٩٩١). *Differential Item Functioning: performance by sex on reading comprehension tests*. ERIC Document ED ٣٣١٨٤٤. Rockville, MD: Educational Resources Information Center.

Geranpayeh, A., & Kunnan, A.J. (٢٠٠٧) Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly, ٤*(٢): ١٩٠-٢٢٢.

Hambleton, R.K., & Rogers, H.J. (١٩٨٩). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, ٢*, ٣١٣-٣٣٤.

Hidalgo, M. H., López-Pina, J. A. (٢٠٠٤). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, ٦٤*, ٩٠٣-٩١٥.

Holland, P. W., & Thayer, D.T. (١٩٨٨). Differential item functioning detection and the Mantel-Haenszel procedure. In: Wainer, H. and Braun, H.I (eds) *Test Validity* .Hillsdale, NJ: Elbaum, pp.١٢٩-١٤٥.

Jiao, H., & Chen, Y. (٢٠١٤) Differential item and testlet functioning analysis. In: Kunnan AJ (ed)
The Companion to Language Assessment (١st ed.). John Wiley and Sons.

Jodoin, M. C., & Gierl, M.J. (٢٠٠١). Evaluating type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, ١٤*, ٣٢٩-٣٤٩.

Karami, H. (٢٠١١). Detecting gender bias in a language proficiency test. *International Journal of*
Language Studies, ٥(٢): ٢٧-٣٨.

Kim, M. (٢٠٠١). Detecting DIF across the different language groups in a speaking test. *Language*
Testing, ١٨(١):٨٩-١١٤.

Kirk, R.E. (١٩٩٦). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, ٥٦*, ٧٤٦-٧٥٩.

Kunnan, A. J. (٢٠٠٧). Test fairness, test bias and DIF. *Language Assessment Quarterly, ٤(٢):* ١٠٩-١١٢.

Kunnan, A. J. (٢٠١٠) Fairness matters and Toulmin's argument structures. *Language*

ISSN ٢٢٥١-٧٣٢٤

*Testing, ٢٤(٢):*١٨٣-١٨٩.
Lin, J., & Wu, F. (٢٠٠٣). Differential performance by gender in foreign language testing. Poster for the ٢٠٠٣ annual meeting of NCME in Chicago.
Li, H., & Suen, H. K. (٢٠١٣). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing,* ٣٠(٢): ٢٧٣-٢٩٧.
Mielikainen, A. (١٩٨٨). Women as preservers and innovators of spoken language. In: Laitinen L(ed), *Isosuinennainen*. Helsinki: Yliopistopainos, pp.٩٢–١١١.
Pae, T. (٢٠٠٤). Gender effect on reading comprehension with Korean EFL learners. *System,* ٣٢(٢): ٢٦٥–٢٨١
Pae, T., & Park, G.P. (٢٠٠٦). Examining the relationship between differential item functioning and differential test functioning. *Language Testing,* ٢٣(٤):٤٧٥–٤٩٦.
Park , G. P., & French, B. F. (٢٠١٣). Gender differences in the Foreign Language Classroom Anxiety Scale. *System* ٤١(٢):٤٦٢-٤٧١
Penfield, R.D. (٢٠٠٥). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement,* ٢٩, ١٥٠–١.
Penfield, R.D. (٢٠٠٩). DIFAS ٥.٠. User's manual. Manuscript in preparation
Penfield, R.D, & Camilli, G. (٢٠٠٧). Differential item functioning and item bias. In: Sinharay, S. and Rao, C.R. (eds.), *Handbook of Statistics, Volume ٢٦: Psychometrics.* New York: Elsevier, pp.١٢٥-١٦٧.
Rezaee, A., & Shabani, E. (٢٠١٠). Gender differential item functioning analysis of the University
of Tehran English Proficiency Test. *Pazhuhesh-e Zabanha-ye Khareji,* ٥٦, ٨٩-١٠٨.
Roever, C. (٢٠٠٧). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly,* ٤(٢):١٦٥-١٨٩.
Rogers, H. J., & Swaminathan, H. (١٩٩٣). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement,* ١٧, ١٠٥-١١٦.
Ryan, K., & Bachman, L. F. (١٩٩٢). Differential item functioning on two tests of EFL proficiency. *Language Testing,* ٩(١): ١٢-٢٩.
Shimizu, Y., Zumbo, B.D. (٢٠٠٥) .Logistic regression for differential item functioning: A primer.
Japan Language Testing Association Journal, ٧, ١١٠–١٢٤.
Swaminathan, H., & Rogers, H.J. (١٩٩٠). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* ٢٧, ٣٦١-٣٧٠.
Tannen. D. (١٩٩٠). *You just don't understand: women and men in conversation*. New York: William Morrow.
Uiterwijk, H., & Vallen, T. (٢٠٠٥). Linguistic sources of item bias for second generation immigrants in Dutch test. *Language Testing* ٢٢*(٢)*: ٢١١-٢٣٤
Wainer, H., & Lukhele, R. (١٩٩٧). How reliable are TOEFL scores? *Educational and Psychological Measurement,* ٥٧(٥): ٧٤١-٧٥٩.
Xi, X. (٢٠١٠).How do we go about investigating test fairness? *Language Testing,* ٢٧(٢):١٤٧-١٧٠.
Zhang, Y., Dorans, N. J. & Matthews-Lopez, J. L.(٢٠٠٣) .Using DIF dissection method to assess effects of item deletion. *Research Report No.* ٢٠٠٥-١٠. College Board

ISSN ٢٢٥١-٧٣٢٤

Zieky, M. **(**١٩٩٣). Practical questions in the use of DIF statistics in test development. In: Holland
P. & Wand
Wainer, H. (eds) *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum, pp. ٣٢١-٣٣٦.
Zumbo, B. D. (١٩٩٩). *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 Resource Research and Evaluation, Department of National Defense.
Zumbo, B. D., & Thomas, D.R. (١٩٩٧). *A measure of effect size for a model-based approach for studying DIF*.
Working paper of the Edgeworth Laboratory for Quantitative Behavioral Sciences, University of Northern British Columbia: Prince George, B. C.