Journal of English Language
Teaching and Learning
Year 5, No.7

Academic Discipline DIF in an English Language Proficiency Test

Seyyed Mohammad Alavi*
Associate Professor of TEFL, University of Tehran
Abbas Ali Rezaee
Assistant Professor of TEFL, University of Tehran
Seyyed Mohammad Reza Amirian
PhD Candidate of TEFL, University of Tehran

Abstract

The purpose of this study was to detect differentially functioning items in the University of Tehran English Proficiency Test (UTEPT) which is a high stake test of English developed and administered by the Language Testing Centre of the University of Tehran. This paper is based on the answers of 400 test takers to the test. All participants earned a master degree either in humanities or science and engineering. To achieve the purpose of this study, the participants were divided into two equal groups. The results of generalized Mantel-Haenszel indicated that out of 100 items 12 items were displaying DIF. Logistic regression procedure also flagged 14 items as exhibiting DIF. Nevertheless, the associated test of effect size for logistic regression showed that none of these effect sizes were large according to the guidelines proposed by both Zumbo and Thomas (1997) and Jodoin and Gierl (2001). Therefore, it was concluded that UTEPT doesn't show significant academic discipline DIF and is equally fair to both humanities and science and engineering groups.

Key words: DIF, Generalized Mantel-Haenszel, Logistic Regression, DIF magnitude.

-تاریخ وصول : ۹۰/۶/۱۷ تأیید نهایی: ۹۰/۸/۱۷

*-Email:Smalavi@ut.ac.ir

Introduction

Detecting Differential Item Functioning (DIF) is essential in all tests in general and in language tests in particular in which test-takers have diverse backgrounds, because DIF items pose a considerable threat to the validity of the test (Kim, 2001). DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure (Zumbo, 1999). In other words, DIF is said to be present when the probabilities of success on a given item are variant between the two groups at the same ability level. A DIF item may be considered biased when a score difference between two or more groups is due to a factor that is not the construct being tested.

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, 1993). DIF is a necessary but not sufficient condition for bias. It is suggested that bias only exists if the difference is illegitimate, i.e., if both groups should be performing equally well on the item.

An item may show DIF but not be biased if the difference is due to actual differences in the groups' ability needed to answer the item, for example if one group is high proficient and the other is low , the low proficient group would necessarily score much lower. DIF can be viewed as bias if the difference is caused by construct-irrelevant factors. In such cases, the item measures another construct, in addition to the one it is supposed to measure. In other words, an item that shows DIF needs to be investigated further to uncover the reasons for its differential functioning. Most DIF analyses compute DIF for a potentially disadvantaged group which is also known as the

focal group, compared to the potentially advantaged group which is also known as the reference group.

Two types of DIF i.e., uniform and nonuniform DIF, are proposed in the literature. Uniform or unidirectional DIF exists when the probability of endorsing an item is greater for one group than for the other group over all the levels of proficiency. In other words, uniform DIF occurs when there is no interaction between the ability level and group membership. On the other hand, non-uniform or crossing DIF exists when 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.

Various methods have been adopted for DIF detection in the literature. One of the most popular ones that is used as the primary DIF detection device at the Educational Testing Service (ETS) is Mantel-Haenszel procedure. Logistic regression, also, has been widely used to detect DIF. An account of these methods appears below.

1.1 Mantel-Haenszel

In their seminal paper, Mantel and Haenszel (1959) introduced a new procedure for the study of the performances of matched groups. Holland and Thayer (1988) adapted the procedure for use in assessing DIF. Mantel-Haenszel (MH) DIF plays an important role in the assessment of the appropriateness of test forms intended for administration to examinees from populations that contain identifiable disjoint subpopulations, such as ethnic groups, men and women, or socio-demographically or geogrographically identifiable groups (Longford, Holland, & Thayer, 1993). MH is a member of contingency table-related approaches which is based on contingency tables and observed conditioning variable. It is a nonparametric approach for identifying DIF which explicitly matches the examinees from two different groups on the ability of

interest, and then compares the likelihood of success on the item for the two groups across the score scale.

Rogers and Swaminathan (1993) consider the MH procedure as one of the most popular procedures for detecting DIF. The primary reasons for its popularity are suggested as its computational simplicity, ease of implementation, and associated test of significance. Furthermore, it can be used with fewer examinees, is easy to program and is easy in terms of computer time (Fidalgo, Mellenbergh, and Muniz, 2000). However, these advantages are obtained at the cost of some generality. The MH was designed to detect uniform DIF and may not be appropriate for detecting non-uniform DIF.

1.1 Generalized Mantel-Haenszel

Generalized Mantel-Haenszel (GMH) statistic, an extension of MH method, is a unified framework for the analysis of DIF using the MH methods. From the outset, various extensions have been proposed for MH statistics, all of them being particular cases of sets of contingency tables. In the case of GMH statistic, the H_0 of no association will be tested against different alternative hypotheses (H_1) that will be a function of the scale on which the factor and the response are measured.

Due to advantages and efficiency of GMH procedure for DIF detection, it has been used as the first method to investigate DIF in the performances of participants on a 100 item test battery. The test battery will be described below. Nevertheless, since GMH doesn't provide any information on the effect size of DIF nor on the direction of DIF, the Logistic Regression (LR) procedure is also used in this study to complement the GMH method. Besides, it would enable us to examine how comparable the results of these two DIF detection procedures are.

1.2 Logistic Regression

The second method adopted for DIF detection in this study is Logistic Regression (LR). LR for detecting DIF was first proposed by Swaminathan & Rogers (1990). Wiberg (2007, p.13) argues that "LR for detecting DIF is based on the probability of answering an item correctly by group membership and conditioning variable." Recently, LR has widely been used to detect DIF because, unlike MH method that can only detect Uniform DIF, LR method is capable of simultaneously detecting both Uniform and Nonuniform DIF. LR is nonparametric, it can be applied to dichotomous and rated items, and requires less complicated computing than IRT-based analyses. Zumbo (1999) suggests an examination of LR using classical statistical software such as Statistical Package for Social Sciences (SPSS). LR is similar to the other DIF detection techniques in that it focuses on DIF at the item level and is mostly applied to dichotomous items. It assesses to what extent item scores can be predicted from total scores alone, from total scores and group membership, or from total scores, group membership and interaction between total scores and group membership.

In LR equation, if proficiency differences by themselves are sufficient for predicting scores with very little residual variance, there is probably little or no DIF in the data. However, if proficiency differences alone do not predict score well and leave a large residual, and, when group membership is added to the equation, predictions become much more precise, then there is DIF present based on group membership. This is uniform DIF because test takers at any score level would be equally affected. In cases, where total score and group membership still do not clear up the residual variance, the interaction term is added, which should make the model more suitable and indicates non-uniform DIF. LR is essentially a model comparison

procedure because it creates and then compares three regression models.

The purpose of the present study is to investigate the comparability of items in UTEPT for examinees of different academic backgrounds, i. e. humanities and science and engineering, as reference and focal groups, respectively. More specifically, this study uses GMH and LR statistics to assess the performance of examinees in humanities and science and engineering language groups on the test after they are matched on English language ability as defined in this study. To this aim, the present study will make use of GMH and LR to conduct a DIF study on UTEPT test. Moreover, the magnitude of detected DIF indexes (effect size) and the direction of DIF are computed through LR procedure.

2.DIF in Language Testing

Understanding and accounting for DIF has become a particular concern for educational researchers to ensure test fairness for all examinees. However, in the context of foreign language proficiency testing group differences have only been explored to a limited degree. Hence, some of the relevant empirical literature on DIF is reviewed below.

Ryan and Bachman (1992) investigated differential performance on the TOEFL (the Test of English as a Foreign Language) and the FCE (the First Certificate of English). Ryan and Bachman found little evidence that males and females performed differently at the item level on either test.

In another study, Hyde and Linn (1997) conducted a meta-analytical study examining gender differences in verbal ability. Generally the meta-analysis demonstrated no significant differences in vocabulary, although there was significant heterogeneity in the effect size. In terms of

reading comprehension, five out of the 21 studies reported a significant difference in favor of males, while ten found significant differences in favor of females. Generally, females were found to have slight disadvantages in reading, speaking, writing, and general verbal ability, but the differences were so small that Hyde and Linn argued that gender differences in verbal ability no longer existed. In contrast, in a comprehensive study conducted in ETS (Cole, 1997) completely different results have been found. In this study in which 400 tests and millions of applicants were investigated, it was reported that a language advantage for females had remained unchanged compared with 30 years ago. Female superiority in verbal ability ranged from noticeable differences in writing and language use to very small differences in reading and vocabulary reasoning.

In a rather recent study, Lin and Wu (2003), examined an English Proficiency Test in China using SIBTEST. They employed SIBTEST for DIF/DBF analyses and DIMTEST for dimensionality testing. The results indicated that although English Proficiency Test did not demonstrate much gender DIF, The SIBTEST and DIMTEST analyses identified and confirmed the presence of the bundle of listening comprehension obviously favoring females, and the bundles of grammar and vocabulary, and cloze favoring males slightly.

2.1 Academic Background DIF

Among very few studies conducted on academic background DIF, Pae (2004) used Item Response Theory (IRT) Likelihood Ratio (LR) approach to investigate DIF in the English subtest of the 1998 Korean National Entrance Exam for Colleges and Universities for examinees with humanities and science backgrounds. The English subtest consisted of a total of 55 items, which comprised the

subscales of listening comprehension and reading comprehension. Data were fitted with a modified three parameter logistic IRT model, and DIF was detected using both the Mantel-Haenszel procedure and the IRT likelihood ratio approach. To evaluate a baseline estimate of DIF due to chance, Pae (2004) repeated the entire DIF detection procedure with the sample randomly divided in half. Pae found 18 DIF items with 28 DIF parameters out of a total of 55 test items across the two subscales at alpha level of 0.05. By academic group, seven items were easier for the sciences. All 12 items identified as exhibiting nonuniform DIF across the two subscales were more discriminating for examinees in the humanities track and covered various topics. This finding suggests that there appeared to be no systematic relationship between item content and DIF directions in terms of item discrimination. In regard to item difficulty, however, seven items were easier for the humanities and nine items were easier for the sciences. For the listening comprehension subscale, items favoring the sciences dealt with number counting and a job interview, whereas items favoring the humanities mostly concerned human relationships. With regards to reading comprehension subscale, seven items were easier for the humanities, whereas four items were more difficult for the sciences.

Since the literature on academic background DIF is so meager, the present study is an effort to investigate the presence of academic discipline DIF in a proficiency test in a foreign language context in order to flag differentially functioning items and identify the potential sources of bias that might be a threat to the validity of the test.

3.Method

3.1 Participants

The participants of the present study were 400 test takers who took UTEPT test. A cut off score was considered as a prerequisite for these participants to be allowed to take part in their PhD exam of the University of Tehran. Based on their academic background, the participants were divided into a reference group (N= 200) with humanities background and a focal group (N=200) with science and engineering background. The humanities group consisted of applicants of social sciences, law, political sciences, management, Persian literature, and foreign languages, and the science and engineering group is comprised of students of chemistry, physics, mathematics, biology, agricultural engineering, mechanical engineering, electrical engineering, and civil engineering. There are both male and female test takers in the sample.

3.2 Instrumentation

The performances of participants on a version of University of Tehran English Proficiency Test (UTEPT) were used in this study. UTEPT is a high stake test of English developed and administered by Language Testing Centre (LTC) of University of Tehran (UT). It is a prerequisite for master's degree holders aiming at participating in PhD exams of UT. It consists of 100 items that contains three sections:1) structure including 15 multiple choice items, 10 written expression items and 5 grammar in context, 2) vocabulary consisting of 30 multiple choice items and 5 vocabulary in context, and 3) reading comprehension comprising 26 reading comprehension and 4 restatement items. This section includes 6 passages each followed by 4 to 8 questions.

There are both short and long passages in reading section with a range of 95 to 359 words.

4.Analysis

4.1 Generalized Mantel-Haenszel

For GMH DIF detection, all 400 scores were entered into the GMHDIF program (Fidalgo, 2010). This program permits, through a single significant test, simultaneous evaluation of DIF in several groups. It is applicable to both dichotomous and polytomous items. In this study, the items are dichotomous. The generalized MH statistic computed by the program for DIF detection tests the null hypothesis of no difference among all the groups. In other words, with a single test we can determine whether an item is free of DIF. In the present study, there were only two groups but the program is able to compare multiple groups simultaneously.

To run GMHDIF program, in the first step the humanities group (N= 200) was selected as the reference group and was dummy codded as 1, and the science and engineering group (N=200) was selected as the focal group and was dummy coded as 2. Then, all 100 items were used as the matching variable and the two- stage $Q_{GMH\ (1)}$ was run. Alpha level was set at $P<0.05$. In the first stage, the program detects DIF items and removes these items from the matching criterion for a second analysis in stage 2. Then, $Q_{\ GMH\ (1)}$ is computed for both stages and DIF items are marked by an asterisk in the output.

The GMHDIF implements two generalized Mantel-Haenszel statistics. They are $Q_{GMH\ (1)}$ or the generalized nominal MH statistic and $Q_{GMH\ (2)}$ or the generalized ordinal MH statistic. Since the data for the

present study is binary, Generalized Nominal MH statistic ($Q_{GMH\ (1)}$) was applied for data analysis purposes.

To summarize, to use GMHDIF program for DIF analyses the following steps were taken:

1. The data was imported for the analysis.
2. Information was provided about the following variables:
   A. Items to be studied for DIF.
   B. Items to be used as the matching variable.
   C. Grouping variable.
3. The desired generalized MH statistic was selected: that is $Q_{GMH\ (1)}$.
4. The results of the DIF analyses were examined.

4.2 Logistic Regression (LR)

To use LR for academic DIF analysis, the SPSS Nagelkerke syntax for nominal data written by Zumbo (1999) was used to analyze each item individually. This syntax could be used for DIF detection in both ordinal and nominal data. Since UTEPT questions are in the form of multiple choice items which are converted to wrong and right responses, Nagelkerke binary syntax was employed for the present study.

At First, the total score of test takers on UTEPT was entered into the equation, and the program computed residuals between the expected item score and the actual item score across all the items for the reference and focal groups. If the expected score turn out to be significantly different from the actual score, it is indicated that total score is not a good predictor of the item score and the group term is added to the regression equation to see how the model changes. And finally the interaction between the group and the score is entered into the regression equation.

The logistic regression procedure uses the item response (0) for incorrect response or (1) for correct response, as the dependent variable, with grouping variable (dummy coded as 1=reference, 2=focal), total scale score for each subject (characterized as variable TOT) and a group by TOT interaction as independent variables. This appears in equation (1). This method provides a test of DIF conditionally on the relationship between the item response and the total scale score, testing the effects of group for uniform DIF, and the interaction of group and TOT to test non-uniform DIF.

Equation (1)Y = b0 + b1 TOT + b2 Discipline+ b3 TOT * Discipline.

Moreover, to see how large the magnitude of DIF is, R-squared, a weighted least squares effect size measure used in LR DIF detection, was computed. The guidelines proposed by Zumbo and Thomas (1997) were used to quantify the magnitude of uniform or nonuniform DIF in this study. This effect size was obtained by comparing the R-squared value of the model in step 3 with that of the model in step 1. The guidelines of Zumbo and Thomas are as follows:

- Type A items- negligible DIF: $R^2 < 0.13$
- Type B items- moderate DIF: $0.13 \leq R^2 \leq 0.26$
- Type C items- large DIF: $R^2 > 0.26$.

It is noteworthy that the latter two categories require that the item be flagged as being statistically significant with two degree of freedom Chi-squared test.

Jodoin and Gierl (2001) proposed a more conservative classification criteria. These criteria were based on the SIBTEST effect size measure (Roussos & Stout, 1996) as a predictor of R-

squared. Accordingly, Jodoin and Gierl (2001) used the following criteria:

- Type A items: negligible DIF: $R^2 < 0.035$
- Type B items – moderate DIF: $0.035 \leq R^2 \leq 0.070$,
- Type C items-large DIF: $R^2 > 0.070$,

In the current study both guidelines were employed to examine the magnitude of obtained DIF sizes and compare the results of two criteria.

5.Results

5.1 Descriptive Statistics

The descriptive statistics for the reference group (humanities) and the focal group (science and engineering) are presented in Table 1. There are 200 humanities with a mean of 50.63 and standard deviation of 14.98 and 200 science and engineering test takers with a mean of 53.31 and standard deviation of 12.14. The table shows that science and engineering group slightly outperformed humanities. Moreover, their scores are more homogenous than those of humanities.

Table 1. Descriptive statistics

MAJOR	N	Mean	Std. Deviation	Std. Error Mean
Humanities	200	50.63	14.9	1.05
Science and engineering	200	53.31	12.14	0.85

As indicated in Table 1, the difference between mean scores of humanities and science and engineering groups is not large. To examine the significance of the difference between the mean score of the two groups, the mean scores were subjected to an independent-sample t-test. The results show that the proficiency levels of humanities and science and engineering groups are not significantly different.

5.2 GMH Results

DIF detection using GMHDIF involves three phases. In phase 1, an initial analysis is performed having eliminated from the matching variable all those items that had been detected with DIF. Then, using the purified variable obtained in phase one items are analyzed. Finally, the same generalized MH statistics is applied to the items detected with DIF in phase two, but this time comparing the groups two-by-two using a Bonferroni-adjusted alpha level (Fidalgo and Scalon, 2010).

The results of the study indicated that out of a total of 100 items, 12 items were flagged as displaying DIF through GMH procedure. 8 of these 12 items showed DIF at both stages and 4 items displayed DIF only at one stage. DIF items include item 6 which is a multiple choice grammar item. In terms of the vocabulary section, items 38, 40, 43, 47, 50, 52, and 58 were found to display DIF. Finally, with regard to the reading section it was revealed that items 67, 73, and 75 exhibited DIF. Table 3 summarizes the results of GMHDIF.

Table 3. Results of generalized Mantel-Haenszel DIF

ITEM No.	QMH stage 1	P	QMH Stage 2	P
6	7.50	0.00*	7.42	0.00*
38	2.75	0.09	5.96	0.01*
40	8.25	0.00*	8.83	0.00*
43	6.37	0.01*	6.20	0.01*
47	6.60	0.01*	6.73	0.00*
50	4.49	0.03*	3.70	0.05
52	2.68	0.10	8.26	0.00*
55	9.13	0.00*	13.1	0.00*
58	3.91	0.04*	4.39	0.03*
67	4.31	0.03*	4.04	0.04*
73	4.52	0.03*	3.21	0.07
75	4.98	0.02*	5.25	0.02*

*the p value is significant at $p<0.05$

Since GMHDIF program is not able to provide us with the magnitude of DIF size, and also in order to improve the reliability of the study, we used LR as a second method of DIF detection that allows us both to examine the significance of DIF size and to test the effect size of DIF.

5.3 Logistic Regression Results

To detect DIF, LR compares three hierarchical regression models and tests whether or not the independent variable entered the equation at each step contributes to the model fit to the data significantly. Out of a total of 100 items, 14 items are flagged as DIF items based on the two-degree-of-freedom chi-squared test of DIF. As indicated in the last column of Table 4, the obtained p values are smaller than 0.05 for all DIF items which shows that these items are displaying significant DIF size.

The R-squared for the 2 degrees-of-freedom DIF test can be found in the fourth column which is computed from the difference between R-squared values of step 3 and step 1. Moreover, in column three the R-squared at step 2 is compared to that of 3 to see how much adding the non-uniform DIF variable contributes to the model. As indicated in Table 4, the difference in R-squared from step 2 to step 3 is small in most of the items suggesting that DIF is predominantly of uniform nature. While items 6, 17, 39, 40, 52, 55, 58, 86, and 92 exhibited uniform DIF, there were only 5 items that displayed non-uniform DIF, i.e., items 35, 37, 41, 47, and 81. Therefore, it is concluded that DIF is predominantly of uniform nature in our data. In other words, it is the grouping variable that is the main source of DIF not the interaction effect.

Table 4. Chi-squared test of DIF; uniform, nonuniformand DIF R^2

Item No.	step2-step1 (uniform DIF R^2)	step3-step2 (nonuniform DIF R^2)	step3-step1 DIF size R^2	χ^2 test of DIF	p
6	0.018	0.003	.021	6.738	.034
17	0.014	0.007	.021	7.405	.025
35	0.002	0.0 27	.029	9.167	.010
37	0.008	0 .017	.025	7.800	.020
39	0.021	0.000	.021	6.206	.045
40	0.021	0.002	.023	7.082	.029
41	0.000	0.029	.029	8.707	.013
47	0.014	0.046	.060	18.741	.000
52	0.018	0.005	.023	7.576	.023
55	0.023	0.006	.029	9.485	.009
58	0.025	0.005	.030	9.513	.009
81	0.000	0.022	.022	6.497	.039
86	0.018	0 .004	.022	6.885	.032
92	0.014	0.004	.018	6.487	.039

Among 14 logistic regressions DIF items observed, 3 items belonged to the grammar section including item 6 that is a multiple choice item aiming to test word order. Item 17 a written expression item, and item 35 a grammar in context item. Moreover, 8 vocabulary items were marked as DIF items which contained items 37, 39, 40, 41, 47, 52, 55, and 58. Taking the reading section into consideration, three items demonstrated considerable DIF

values , i.e. items 81 (passage three on history), 86 (passage four on theories in education), and 92 (passage six on painting art). These items are expected to test main ideas of the passages.

As it was pointed out earlier and emphasized by many researchers, in statistical hypothesis testing the test statistic should be accompanied by some measure of the magnitude of the effect because "the effect size prevents flagging unimportant differences in large samples" (Monahan, McHorney, Stump, and Perkins; 2007, p.104). This is necessary because small sample sizes can hide interesting statistical effects whereas large sample sizes can point to statistically significant findings where the effect is quite small and meaningless. Zumbo and Thomas (1997) indicate that an examination of both the 2-df Chi-square test (of the likelihood ratio statistics) in logistic regression and a measure of effect size is needed to identify DIF.

As indicated in Table 4, the effect size of all DIF items is smaller than 0.13 and can be considered as type A or negligible effect size based on Zumbo and Thomas' (1997) guidelines. However, if we apply Jodoin and Gierl's (2001) more conservative criteria, item 47, for which the obtained R-squared is 0.06, shows moderate DIF and is identified as type B item. Apart from item 47, the obtained R-squared values of all other items are negligible based on Jodoin and Gierl's criteria. This corroborates the findings of Hidalgo and Lopez-Pina (2004) who contended that when the criteria adopted by Jodoin and Gierl were used, a slightly larger percentage was classified as having moderate DIF. Hence, it can be concluded that the magnitude of DIF is not statistically significant for any of DIF items. In other words, UTPET items do not function differentially across two groups. It means that both humanities and science and engineering groups have an equal chance of excelling on UTEPT.

With regard to the direction of DIF, it was found that 5 uniform DIF items, i.e., items 6, 17, 40, 58, and 92- favored humanities group while 4 other uniform DIF items, i.e., items 39, 52, 55, and 81- favored science and engineering group. As far as nonuniform DIF is concerned, items 35, 37, 41, and 47 were differentially easier for the science and engineering and only item 81 was easier for humanities.

Item 47, the only item with moderate DIF size flagged in the study, is a multiple choice vocabulary item where test takers have to choose the best synonym for the word "tangible". The correct response is "concrete". As mentioned in the previous paragraph this item favors science and engineering group. This might be due to the fact that the words "tangible" and "concrete" are more frequently found in scientific texts and examinees in science and engineering group are more likely to come across these words in their textbooks.

6.Discussion

The results of GMH indicated that out of 100 items, 12 items displayed DIF. Through a single test, GMH statistics permits a simultaneous evaluation of DIF in several groups, being applicable to both dichotomous and polytomous items. However, since there is no test of magnitude, that is DIF effect size, associated with GMHDIF, we cannot confidently assert that UTEPT functions differentially for humanities and science and engineering groups based on our GMH findings. Thus, the results of LR could be more illuminating in this case.

The logistic regression findings revealed that out of a total of 100 items, 14 items exhibited DIF. Since the test of significance alone is not sufficient to flag an item as biased item and it is recommended in the literature to examine the

results in the context of some measure of effect size (Kirk, 1996), the combination of statistical test with an effect size measure was used to reduce false identification rates.

In general, the results of logistic regression suggested that the effect size of DIF is negligible for almost all items. This finding is consistent with the findings of Rezaee and Shabani (2010) who found that none of UTEPT DIF items demonstrated a moderate or large DIF effect size. Nevertheless, in this study, item 47, which is a vocabulary item, exhibited moderate size or type B DIF. To determine if the flagged DIF items exhibit uniform or nonuniform DIF, the gained R-squares were studied. According to Zumbo (2001), uniform DIF exists when there is no interaction between ability level and group membership and nonuniform DIF occurs when there is an interaction between ability level and group membership. Therefore, based on the obtained R-squares, it turned out that 9 items exhibited uniform DIF, and 5 items displayed non-uniform DIF. In other words, for 9 items group membership alone accounts for the observed DIF size and for 5 items the interaction between ability level and group membership explains the gained DIF value.

To determine the direction of DIF, the value of the grouping factor was examined. A negative value signifies that the item favors the reference group, i. e. humanities, and a positive value indicates that the focal group, i. e. science and engineering, group is favored. Out of a total of 14 DIF items, 6 items favored the reference group that is humanities group. On the other hand, 8 items also were in favor of the science and engineering group. It is interesting that, contrary to our expectation, in this study more items were in favor of the focal group rather than the reference group.

The results of the study suggest that LR and the generalized MH procedures are rather comparable. 5 items

were identified as DIF items by both procedures that is items 6, 47, 52, 55, and 58. LR generally detected more DIF items (14 items) than the generalized MH procedure (12 items). This is in line with the findings of Muniz, Hambleton, and Xing (2001) who found that Mantel-Haenszel procedure does not show its strength unless group sizes are large. Only with reference and focal groups of 500 did Mantel-Haenszel outperform the other procedures. With small groups, it often performed worse than other procedures. It also reinforces the findings of Hidalgo and Lopez-Pina (2004) who suggested that LR analysis generally detected more items with DIF than MH procedure.

By academic discipline, 9 DIF items identified as displaying uniform DIF were easier for humanities while 5 nonuniform DIF items were more discriminating for the examines in the science group. All in all, this finding suggests that there is no systematic relationship between group membership and performance on UTEPT. In other words, UTEPT items do not differentially favor humanities or science and engineering groups and are not discriminating for examinees in either group.

The relationship between DIF direction and item content also deserves attention. The results indicated that 8 identified DIF items were easier for the science and engineering group and 6 items were easier for humanities group. Considering the content of DIF items, we find out that items favoring the humanities group belonged to all three sections of the test, i.e. structure, vocabulary and reading comprehension. The same story is true about the items that were more discriminating for the science and engineering group. Therefore, no systematic relationship could be discerned between item content and DIF directions in terms of item difficulty.

7.Conclusions and Implications

The purpose of the present study was to study UTEPT for DIF. Out of a total of 100 items, 12 items were flagged by generalized Mantel-Haenszel procedure. One of the drawbacks of generalized Mantel-Haenszel method is that it cannot calculate the magnitude of DIF size. Consequently, logistic regression was also employed as a second method for DIF investigation. 14 items were detected as DIF items by logistic regression method. Moreover, further analysis of DIF items revealed that 6 items were differentially easier for humanities, and 8 items were easier for the science and engineering group. Taking DIF effect size into account, however, none of DIF items exhibited sizable DIF magnitude, signaling that the actual DIF effect size was negligible. In other words, UTEPT doesn't function differentially across the two groups. Hence, the developers of UTPET could be confident that the test is not biased and a construct irrelevant factor such as academic group doesn't harm the validity of the test.

The result of the study revealed that Zumbo-Thomas (1999) and Jodoin and Gierl (2001) tests of effect size are rather compatible. This finding reinforces Gierl and his colleagues' findings who showed that the Zumbo-Thomas effect size measure is correlated with other DIF techniques like the MH and SIBTEST hence lending validity to the method. It also aligns with the findings of Zheng, Gierl, and Cui (2005) study who discovered a high correlation among DIF effect size measures.

With regard to the content of DIF items, it was revealed that no systematic relationship exists between the direction of DIF and the content and section (vocabulary, grammar, or reading comprehension) of the DIF item.

It should be noted that in the current study the researcher made use of the data obtained from only 400 test takers that could be considered as a potential limitation of the study. It

might have affected the results of the study since MH procedure doesn't show its strength unless the sample is large. Therefore, future studies could address the issue of academic discipline DIF using a larger data set. Examining UTEPT for DIF using other methods such as IRT and SIBTEST would be an interesting future study because it would provide further information on the source of DIF items. Moreover, a replication of this DIF study with a more comprehensive content analysis of DIF items would shed more light on the underlying sources of DIF in UTEPT items.

References

-Angoff, W. H. (1993), Perspectives on differential item functioning methodology. In P. W. Holland and H. Wainer (Eds). Differential Item Functioning. New York: Routledge.

-Fidalgo, A. M. Mellenbergh, G. J. & Muniz, J (2000), Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5, 3, 43-53.*

-Fidalgo, A. M. (2010),User's manual for GMHDIF program. University of Oviedo.

-Fidalgo, A. M. & Scalon, J. D. (2010), Using generalized Mantel-Haenszel statistics to assess DIF among multiple groups. *Journal of Psychological Assessment. 28, 1,60-69.*

-Hidalgo, M.D, & Lopez-Pina, J.A. (2004), DIF detection and effect size: A comparison between Logistic Regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64, 903-915*.

-Holland, P. W. & Thayer, D. T. (1988), Differential item functioning detection and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.). *Test Validity* (pp.129-145) Hillsdale, NJ: Elbaum.

-Jodoin, M. C. & Gierl, M. J.(2001), Evaluating type I error and power rates using an effect size measure with logistic

regression procedure for DIF detection. *Applied Measurement in Education, 14, 329-349*.

-Kim, M (2001), Detecting DIF across the different language groups in a speaking test. *Language Testing, 18, 89-114.*

-Kirk, R. E. (1996), Practical significance: a concept whose time has come. *Educational and Psychological Measurment, 56, 746-59.*

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

-Longford, N. T., Holland, P. W. & Thayer, D. T (1993), Stability of MH D-DIF statistics across populations. In P. W. Holland and H. Wainer (Eds). *Differential Item Functioning*. New York: Routledge.

-Mantel, N. & Haenszel, W. (1959), Statistical aspects of the analysis of data from retrospective studies of diseases. *Journal of the National Cancer Institute, 22 719-748.*

-Muniz, J. Hambleton, R. K. & Xing, D. (2001), Small sample studies to detect flaw in item translations. *International Journal of Testing, 1, 115-135.*

-Pae, T.E (2004), DIF for examinee with different academic backgrounds. *Language testing. 21, 53-73.*

-Rezaee, A. & Shabani, E. (2010), Gender differential item functioning analysis of the University of Tehran English

Proficiency Test. *Pazhuhesh-e Zabanha-ye Khareji,56, 89-108.*

-Rogers, H. J. & Swaminathan, H. (1993), A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17, 105-116.*

-Roussos, L. A. & Stout, W. F. (1996), Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33, 215-230.*

-Ryan, K. & Bachman, L.F. (1992), differential item functioning on two tests of EFL proficiency. *Language Testing, 9, 1 12-29*.

-SPSS Incorporated (2009).SPSS 18. Chicago, IL: SPSS, INC.

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

-Wiberg, M (2007), Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods. UMEA University.

-Zheng, Y. Gierl, M. J. & Cui, Y. (2005), Using real data to compare DIF detection and effect size measures amon Mantel-Haenszel, SIBTEST, and Logistic Regression

Procedures. Center for Research in Applied Measurement and Evaluation. University of Alberta.

-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 Resource Research and Evaluation, Department of National Defence.

-Zumbo, B.D. (2001), Investigating DIF by statistical modeling of the probability of endorsing an item: logistic regression and extensions thereof. Paper presented at the National Council for Measurement in Education meeting, April 2001.

-Zumbo, B. D. & and Thomas, D. R. (1997), A measure of effect size for a model-based approach for studying *DIF*. Working paper of the Edgeworth Labaratory for Quantitative Behavioral Sciences, University of Northern British Columbia: Prince George, B. C.