Applications of Language Studies
jals.hmu.ac.ir
Hazrat-e Masoumeh University

Item Response Theory and Mantel-Haenszel Procedures in Detecting Academic Discipline Differential Item Functioning and Differential Skill Functioning with English Proficiency Test

Hamed Ghaemi[1] Mahsa Khorami[2]

1. Assistant Professor, Bahar Institute of Higher Education, Mashhad, Iran.
ghaemiacademy@gmail.com (corresponding author)

2. Instructor, Bahar Institute of Higher Education, Mashhad, Iran.
mahsa_m6669@yahoo.com

Abstract

The present mixed-method study aimed at investigating the presence of Differential Item Functioning (DIF) and Differential Skill Functioning (DSF) in a high-stakes language proficiency test in Iran, the English Proficiency Test (EPT) with different academic backgrounds (i.e., Humanities vs. Sciences & Engineering) using Item Response Theory (IRT) and Mantel-Haenszel (MH) approaches. It also aimed at detecting if there is any correlation between IRT and MH methods and also detecting which DIF items are biased. The English subtest consisted of a total of 100 items. The participants (N = 642) were selected by convenience sampling from universities in Tehran. The results displayed DIF between Sciences and Humanities students, but they did not show DSF in favor of a particular academic discipline group. Hence, on the basis of the findings, it was concluded that the EPT scores are not free of construct-irrelevant variance but because 14 DIF detected items out of 100 items were too small, the overall fairness of the test was confirmed. In addition, it was found that some of DIF detected items were bias and some of them functioned differently, simply because the two groups differed in their abilities. The positive correlation between IRT and MH methods was also proved.

Keywords: Differential Item Functioning, Differential Skill Functioning, Item Response Theory, Mantel-Haenszel, English Proficiency Test

Citation: Ghaemi, H., & Khorami, M. (2024). *Applications of Language Studies*, *2*(2), 20-58.

Received: 2024-02-07
Revised: 2024-04-04
Accepted: 2024-07-20
Article type: Research paper

P-ISSN: 2588-6975
E-ISSN: 2588-6983
Publisher: Hazrat-e Masoumeh University Press
DOI: 10.22034/jals.2024.2022383.1009

1. Introduction

Reliability and validity are crucial goals in test formulation and design. The degree to which a test measures what it is supposed to assess is referred to as test validity. Validity is a crucial concept because without it, any measurement, inference, or conclusion based on it is useless. It is a truism, according to Geranpayeh and Kunnan (2007), that tests must be fair to test takers in order for test-score interpretations made by test users (e.g., admission officers and employers) to be valid. Quality management in test design, administration, and scoring, adequate coverage of relevant content, sufficient construct validation work, equal learning opportunities and access to testing, and items measuring only the skill or ability under investigation without being unduly influenced by construct-irrelevant variance introduced through test-taker bac, are all examples of test fairness (Saville, 2005; Shohamy, 2000).

When standardized English-language examinations are given to people all over the world, the test-taking population may differ based on personal and educational criteria including age, gender, first language, and academic discipline. Test developers must constantly check their tests to ensure that all test takers are receiving a fair test (Geranpayeh & Kunnan, 2007). Items must be appropriate for readers from a variety of racial, ethnic, cultural, and linguistic backgrounds in order for the exam to be valid (Abedi et al., 2005). So, test validation through various ways is a vital step for fair assessment of examinees as well as proper interpretation of a test, especially when considering the personal and social repercussions of a test.

The process of validating a test begins at the development stage and continues after it has been used. The validation process, according to Zumbo (1999), is never complete. As Bachman (2004) writes:

> Language examinations have become ingrained in both our educational system and our culture. Language test results are used to make assumptions about people's language abilities and to inform decisions we make about them. Language exams, for example, are used to identify second and foreign language learners in schools, to choose students for university entry, to place them in language programs, to screen possible immigrants and to select

> employees. As a result, language exams have the potential to assist us in gathering relevant data that will benefit a wide range of people. However, in order to actualize this promise, we must be able to demonstrate that the scores we acquire from language exams are accurate, as well as the methods we employ to evaluate and use those scores. (p. 3).

Differential Item Functioning (DIF) analysis, according to Li and Suen (2012), is one technique to analyze the appropriateness and validity of tests. It was originally developed to uncover possible item bias across subgroups. "A disparity in item performance between two comparable groups of examinees" is what DIF stands for (Dorans & Holland, 1993). It exists when the chance of successfully answering a question varies between groups (Koo et al., 2013). De Ayala et al. (1999) and Klieme and Baumert (2001) have suggested that multidimensionality may be the origin of DIF, and that DIF happens when one of the unexpected additional dimensions reflected by a score is connected to race/ethnicity, gender, or other demographic characteristics (Li & Suen, 2012).

According to Penfield (2001), the inclusion of DIF items may imply a systematic invalidity of the items, putting one group at a disadvantage. If there is a score discrepancy across groups as a result of unequal item performance, the validity of the scores may be jeopardized (Holland & Wainer, 1993). According to Park and French (2013), the existence of DIF items in an instrument indicates that the score measured by the instrument should be understood and used with caution because it could have various meanings for different group members. As a result, detecting, revising, or removing DIF items from highly selective tests is critical (Yu et al., 2006).

In the context of DIF, many studies have been conducted to identify group differences in performance and to see if test items are invariant across members of different subgroups (e.g., Carlton & Harris, 1992; Chen & Henning, 1985; Elder, 1996; Lawrence & Curley, 1989; Maller, 2001; Pae, 2004; Ryan & Bachman, 1992; Scheuneman & Gerritz, 1990; Shmitt & Dorans, 1990; Thissen et al., 1988, 1993). Ethnicity (e.g., Schmitt & Dorans, 1990), academic backgrounds (e.g., Pae, 2004), linguistic backgrounds (e.g., Chen & Henning, 1985; Ryan & Bachman, 1992), and disability status (e.g., Maller, 1997) have all been used to group people (e.g.Maller, 2001).

DIF analysis has been proposed as a tool for explaining group strengths and shortcomings; Differences in group strengths and weaknesses, on the other hand, rarely arise solely at the item level (Li & Suen, 2012). Li and Suen also explained that even though the item is the DIF's unit of analysis, it is not very useful for studying these differences.

Differential Skill Functioning (DSF) analysis, a variation of traditional DIF methods, has been proposed as an alternative. DSF occurs when examinees from different groups have different probabilities of successful performance in a specific sub-skill underlying the measured construct, despite having the same overall ability (Li & Suen, 2012). According to Li and Suen, although the logic and approach of DSF studies are similar to those of DIF analyses, the aim, or the unit of analysis, is a skill expressed by a number of items rather than a single item. Li and Suen (2012) extended the DIF approach to individual skill performance as evaluated by the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT) in order to compare aggregate groupings to the complete population matched on overall scores (Li & Suen, 2012).

According to McNamara and Roever's (2006) analyses of test and item bias, DIF and DSF are part of the overall work on test fairness. They help test users conduct a test that is fair to all participants and is not biased towards a certain group. If a test is biased, the inferences drawn from it are not based on genuine differences, and any measurement, inferences, or decisions drawn from it are useless. The validity of inferences can be jeopardized by a variety of variables common to language performance evaluations, such as raters, task, rating scale, candidate characteristics, and interactions among these (see McNamara, 1996 for a model and discussion); As a result, researchers and test users should be aware of test bias analysis, how it might affect test results, and how to identify and eliminate it.

Although it has been suggested in literature that language test performance may be influenced by background knowledge (e.g., Clapham, 1996; Hale, 1988), most DIF and DSF studies focus on comparisons between gender (e.g., Carlton & Harris, 1992), ethnicity (e.g., Schmitt & Dorans, 1990), or different language groups (e.g., Ryan & Bachman, 1992).According to Pae (2004), studies that looked at the impact of prior knowledge on test

performance are extremely rare. In this regard, DIF and DSF investigations for test-takers from different academic disciplines (e.g., Humanities vs. Sciences) would fill a gap in the DIF and DSF literature by revealing whether test items and skills function differently for examinees from different academic backgrounds but are matched to the primary construct that the test is intended to measure.

English Proficiency Test (EPT) is a test that is used to examine and assess the competency level of applicants who are interested in taking the Ph.D. exams. Given the importance of validity, test administrators of universities must guarantee that the test is not biased against a specific group. Since few studies have focused on academic background which creates DSF and DIF on different tests (Pae, 2004), test administrators may wonder whether students' academic backgrounds affect their performance or not. Thus, the purpose of the present study is to identify items and skills that exhibit DIF and DSF on EPT for examinees with different academic backgrounds (Humanities versus Sciences) using the Item Response Theory (IRT) and Mantel-Haenszel (MH) (Mantel & Haenszel, 1959) methodologies so that its administrators will be able to improve the validity of it for its applicants. This study examines differential functioning at both the item and skill levels, and if any disparities in item and skill performance are discovered between two groups, it attempts to see whether they can be improved.

2. Research Questions

According to the purpose of this study, the researchers try to answer the following questions:

1. Do items of English Proficiency Test (EPT) show Differential Skill Functioning (DSF) in favor of a particular academic discipline group using Mantel-Haenszel (MH) and Item Response Theory (IRT) procedures?
2. Do items of English Proficiency Test (EPT) show Differential Item Functioning (DIF) in favor of a particular academic discipline group using Mantel-Haenszel (MH) and Item Response Theory (IRT) procedures?
3. Do Item Response Theory (IRT) and Mantel-Haenszel (MH) detect similar Differential Item Functioning (DIF) and Differential Skill Functioning (DSF)?
4. What does the content analysis of DIF items reveal in terms of linguistic differences?

3. Method

3.1. Participants and Setting

The participants of the present study are 642 post-graduate students with an age range of 24 to 38. They were selected by convenience sampling from universities in Tehran. They were going to pursue their PhD studies in a variety of academic disciplines including Humanities, Sciences, and Engineering. Passing EPT is a prerequisite for them to be allowed to sit for the Ph.D. exam. Based on the academic background, the sample is divided into a reference group of 468 participants with Sciences and Engineering background and a focal group of 174 participants with Humanities background. The Sciences and Engineering group comprises students of chemistry, physics, mathematics, biology, mechanical engineering, electrical engineering, and civil engineering; and the Humanities group consists of students of social sciences, law, political sciences, management, Persian literature, and foreign languages. There were both male and female test takers in the sample. Because some studies identified 250 test takers per group as the minimum for adequate power in the MH procedure, a sample size of 642 test takers was chosen (Güler & Penfield, 2009; Mazor et al., 1992; Paek & Wilson, 2011; Rogers & Swaminathan, 1993). When the magnitude of the DIF is large, the power found in simulation studies with smaller sample sizes (50 or 100 test takers per group) is only moderately acceptable (Fidalgo et al., 2004; Fidalgo et al., 2007).

3.2. Materials

3.2.1. EPT

English Proficiency Test (EPT) is a high-stakes test of English that is being developed and administered by English department of the university. It is a prerequisite for master's degree holders aiming at participating in Ph.D. exams. It includes three sections with a time limit of 75 minutes. The first section including structure and written expression questions consists of 30 items, the second part includes vocabulary questions with 30 items, and the last section consists of 60 items including reading comprehension questions.

3.3. Procedures

3.3.1. Data Collection Procedures

The purpose of this study is to examine academic discipline differences in EPT so that its administrator will be able to improve the validity of it for its applicants. This study is part of a large-scale study. As a result, a total of 642 test takers were used for the pursuit of the research questions and data was collected by administering EPT in October, 2016. According to Penfield (2001), comparisons are frequently done between a single reference group, which is the favored group of examinees, and the focal group, which is the disadvantaged group of examinees, while undertaking a Differential Item Functioning (DIF) analysis (Abbott, 2016). As a result, the study's researchers divided the participants into two groups based on their academic discipline: a reference group of 468 students with a Sciences and Engineering background and a focal group of 174 students with a Humanities background, in order to see if the academic discipline of students has an impact on their performance. In DIF and DSF investigations, it is advised that more than one method of DIF or DSF analysis can be used to obtain more reliable results (Camilli, 2006; Camilli & Shepard, 1994; Pae, 2012; Uiterwijk & Vallen, 2005). The current study uses two methodologies to support this claim: (a) the Mantel-Haenszel (MH) model, and (b) the Item Response Theory (IRT) model. This increased the results' dependability and allowed for a comparison of the degree of concordance between the two methodologies' outcomes. The data was analyzed using DIFAS and Winsteps applications in addition to IRT and MH methods.

3.3.2. Mantel-Haenszel Procedures

DIF detection can be done in a variety of ways. The accuracy of them is determined by the disparities in the groups' ability levels. The initial step in detecting DIF is to match examinees with similar skill levels (Allalouf & Abramzon, 2008). Matching is difficult when there are significant differences in ability. Large ability gaps can result in decreased agreement across various DIF detection systems when it comes to DIF items, for example (as was found by Hambleton & Rogers, 1989).

The Mantel-Haenszel (MH) DIF detection method is a widely utilized method (see Holland & Thayer, 1988). Because it has been widely used in a number of studies (e.g., Azocar et al., 2001; Roznowski & Reith, 1999; Scheuneman & Gerritz, 1990; Schmitt et al., 1993), we utilized the Mantel Haenszel chi-square (MH-×2) approach to detect DIF (Mantel

& Haenszel, 1959). The MH technique was chosen for this investigation because of its simplicity, consistency across populations, and ease of interpretation.

When using the MH procedure to detect DIF, examinees are first grouped according to an estimate of ability, usually the total test score, and then a two-by-two contingency table is created for each level of ability, crossing group membership (reference and focal) and item performance (correct and incorrect) (Penfield, 2001).

Finally, in terms of academic discipline directedness, EPT was evaluated for the presence of DIF. Mantel Chi-square statistic (Mantel, 1963; Zwick et al., 1993; Zwick et al., 1997) and Liu-Agresti Cumulative Common Log-Odds Ratio (Liu & Agresti, 1996; Penfield & Algina, 2003) were obtained and analyzed for presence and direction of DIF in favor of reference or focal group using Differential Item Functioning Analysis System (DIFAS) program version 5.0.

3.3.3. Item Response Theory

An alternative approach to the elimination of items could be to investigate if the items that exhibit DIF genuinely assess the same construct across conditions even if they do so using a different method (Makransky & Glas, 2013). Such disparities can be modeled using group-specific item characteristics in Item Response Theory (IRT). This method is only valid if it can be demonstrated that the responses to the items in both groups are related to the same latent variable. To put it another way, the construct being measured must be the same in both groups. This may be demonstrated by examining if the same IRT model applies to the complete set of response data. This technique is justified by the fact that objects can have slightly different true parameters in different contexts. When statistical evidence supports the premise that the items assess the same construct across conditions, these differences can be modeled (Makransky & Glas, 2013). The study used BILOG-item MG's response theory (IRT)-based DIF detection approach, which compares latent trait item difficulty metrics across groups (Geranpayeh & Kunnan, 2007).

The present study was also carried out in the framework of IRT using Winsteps program. The basic premise of IRT is that each test item is defined by one or more

parameters, whereas each test taker is defined by a single ability parameter. The chance that a given test-taker will properly answer a given question is a function of both the item's and the test taker's factors. The response to one item is independent of the responses to other items if those requirements are met (Makransky & Glas, 2013).

4. Results

4.1. Results of Question One

The data was analyzed using DIFAS program to see whether DIF and DSF are present in items of English Proficiency Test across different academic disciplines. With this aim, the participants of Sciences and Engineering disciplines were taken as reference group (468 members) and the participants of Humanities discipline were taken as focal group (174 members). According to Penfield (2003), critical values of statistics of MH CHI and BD columns are 3.84 for a Type I error rate of 0.05 and 6.63 for a Type I error rate of 0.01 and a value greater than 2.0 or less than –2.0 in LOR Z column may be considered as an evidence of the presence of DIF. MH LOR column also has been explored to identify the direction of possible DIF in scale items. Positive values indicate DIF in favor of the reference group (Sciences and Engineering disciplines in this study), and negative values indicate DIF in favor of the focal groups (Humanities discipline). CDR flags any item for which the Mantel-Haenszel chi-square or the Breslow-Day chi-square statistic is significant at a Type I error rate of 0.025, and items that demonstrate DIF are divided into three groups at the Educational Testing Services (ETS) in the United States: Items with a negligible DIF are in Category A, items with an intermediate DIF are in Category B, and items with a large DIF are in Category C (Dodeen & Johanson, 2003). Tables 1, 2, and 3 display flagged items detected by Mantel-Haenszel theory using DIFAS program.

4.1.1. Grammar Skill Using DIFAS Program

As shown in Tables 1 and 2, items of grammar skill are flagged as DIF. Item 23 (MH CHI= 5.95, MH LOR= 0.49, LOR Z= 2.51 and BD= 0.00) has been categorized as having moderate (B) level of DIF and item 26 (MH CHI= 0.09, MH LOR= 0.07, LOR Z= 0.40 and BD= 5.12) has been categorized as having small (A) level of DIF. The value of MH CHI for item 23 is

more than 3.84 and the value of BD for item 26 is more than 3.84 and both of them show DIF toward reference group. In order to prove the existence of DIF items in the MH CHI column, LOR Z column was also studied and it demonstrates item 23 with DIF, because the value of it is more than +2.

Table 1

DIF Statistics: Dichotomous Items of Grammar Skill

NAME	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Q23	5.9533	0.4905	0.1952	2.5128	0.001	Flag	B
Q26	0.0969	0.0751	0.1865	0.4027	5.126	Flag	A

4.1.2. Vocabulary Skill Using DIFAS Program

In Tables 2 and 3, items of vocabulary skill are presented as showing DIF.

Table 2

DIF Statistics: Dichotomous Items of Vocabulary Skill

NAME	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Q38	0.3394	-0.1597	0.2337	-0.6834	9.957	Flag	A
Q41	6.4622	-0.5441	0.2082	-2.6134	0.983	Flag	B
Q49	0.9555	0.1962	0.1855	1.0577	6.136	Flag	A

Item 38 has been categorized as having small (A) level of DIF (MH CHI= 0.33, MH LOR= -0.15, LOR Z= -0.68 and BD= 9.95), item 41 (MH CHI= 6.46, MH LOR= -0.54, LOR Z= -2.61 and BD= 0.98) has been categorized as having moderate (B) level of DIF and Item 49 (MH CHI= 0.95, MH LOR= 0.19, LOR Z= 1.05 and BD= 6.13) has been categorized as having small (A) level of DIF. The value of BD for items 38 and 49 is more than 3.84 and the value of MH CHI for item 41 is more than 3.84. MH LOR column was studied to identify the direction of possible DIF in scale items. Items 38 and 41 show DIF toward focal group, because they have negative values and item 49 shows DIF toward reference group because it has positive value. In order to prove the existence of DIF items in the MH CHI column, LOR

Z column was also studied and it shows that item 41 shows DIF, because its value is less than -2.

4.1.3. Reading Comprehension Skill Using DIFAS Program

Out of 60 items of reading, 5 items are flagged as DIF. Item 64 (MH CHI= 8.20, MH LOR= -0.55, LOR Z= -2.90 and BD= 4.64) has been categorized as having moderate (B) level of DIF, item 68 (MH CHI= 1.54, MH LOR= 0.25, LOR Z= 1.32 and BD= 6.47) has been categorized as having small (A) level of DIF, item 72 (MH CHI= 0.24, MH LOR= 0.13, LOR Z= 0.60 and BD= 7.73) has been categorized as having small (A) level of DIF, item 86 (MH CHI= 0.55, MH LOR= -0.21, LOR Z= -0.87 and BD= 5.84) has been categorized as having small (A) level of DIF and item 90 (MH CHI= 11.51, MH LOR= -0.65, LOR Z= -3.50 and BD= 0) has been categorized as having moderate (B) level of DIF. As shown in Table 3, the value of MH CHI for items 64 and 90 is more than 3.84 and the value of BD for items 68, 72, and 86 is more than 3.84. In MH LOR column, items 64, 86, and 90 have negative values and show DIF toward focal group but items 68 and 72 have positive values and show DIF toward reference group. In order to prove the existence of DIF items in the MH CHI column, LOR Z column was also studied and it shows item 41 with DIF, because the value of it is less than -2. LOR Z column shows 2 items with DIF, both of them were less than -2.

Table 3

DIF Statistics: Dichotomous Items of Reading Skill

NAME	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Q64	8.2091	-0.5598	0.1927	-2.905	4.646	Flag	B
Q68	1.5466	0.2574	0.1946	1.3227	6.471	Flag	A
Q72	0.2468	0.1394	0.2308	0.604	7.732	Flag	A
Q86	0.5549	-0.2146	0.2465	-0.8706	5.841	Flag	A
Q90	11.5106	-0.6513	0.1859	-3.5035	0	Flag	B

Looking back at Tables 1, 2, and 3 and considering that 2 items of grammar skill (items 23 and 26), 3 items of vocabulary skill (items 38, 41, and 49) and 5 items of reading

skill (items 64, 68, 72, 86, and 90) have been flagged as DIF, it is noticed that although the number of DIF items in reading section is more than other skills, this skill does not show DSF, because out of 40 items belonging to reading skill, just 5 items have been flagged as DIF and three of these items indicate that focal group (Humanities academic discipline) is advantaged and two of them indicate that reference group (Sciences and Engineering academic disciplines) is favored. Therefore, it is concluded that items of EPT do not show DSF in favor of a particular academic discipline group using Mantel-Haenszel (MH) procedure.

In order to assess the presence of DIF and DSF by implementing IRT method, Winsteps was used. Two DIF detection methods are used in Winsteps (a Rasch analysis software program): (1) the Welch t-test and (2) the Mantel-Haenszel method (Linacre, 2010). The outcomes of these two methods are used to classify items using the ETS DIF Categories. The interest group is the focal group, whereas the reference group serves as a comparison group (Camilli & Shepard, 1994; Dorans & Holland, 1993). For this study, the participants of Sciences and Engineering disciplines were taken as the reference group (468 members) and the participants of Humanities discipline were taken as the focal group (174 members). Generally, the reference groups are expected to have the perceived advantage while the focal groups have the disadvantage.

Person Class (PC) is the indicator of the group which the participants belong to. PC1 belongs to Sciences and Engineering students (reference group) and PC2 belongs to Humanities students (focal group). DIF measure is the difficulty of item in person class. The hypothesis in the Welch probability column is that each item has the same difficulty for both the reference and focal groups, and it depicts the likelihood of observing the amount of DIF contrast by chance. As a result, values less than 0.05 indicate DIF. Welch t is a Welch's t-statistic that calculates the DIF significance (Runnels, 2013). Both the amount of the difference in logit units between the groups and the statistical significance of the difference should be considered when performing DIF Analysis in a Rasch context (Linacre, 2010). The DIF Contrast measure, according to Runnels (2013), depicts the difference between the item measures, or the difference in item difficulty between the two groups (a difference of at least

0.5 logits is required for the DIF to be noticeable). The DIF Contrast statistic represents the amount and direction of the difference in difficulty estimations between the reference and focal groups. Negative values imply that the item is more difficult for the focal group, whereas positive values suggest that the item is more difficult for the reference group (Runnels, 2013).

The ETS DIF Categories have been in operation for over 25 years and are assigned based on the direction, size, and relevance of the DIF statistics (Linacre, 2010). To "avoid recognizing items that demonstrate practically insignificant but statistically significant DIF," the three categories were devised (Clauser & Mazor, 1998). All items are sorted into the DIF categories as described in Table 4 (Zwick, 2012).

Table 4

ETS Differential Item Functioning Categories from WINSTEPS

ETS DIF Category	Requirements
Class A (Negligible DIF)	|DIF Contrast| < 0.43
Class B (Slight to Moderate DIF)	Doesn't meet the requirements for A or C
Class C (Moderate to Large DIF)	|DIF Contrast|> 0.64

Tables 5, 6, and 7 demonstrate the results of IRT analysis for different academic discipline groups and their performance on different skills of EPT. Note that, due to space limitation, only DIF-flagged items are presented.

4.1.4. Grammar Skill Using Winsteps Program

In Table 5, 3 items are presented as showing DIF detected by IRT method. They are items 8, 20, and 23. Item 8 (DIF contrast= -0.38, Welch t= -2.04 and Welch prob. =0.04) and item 23 (DIF contrast= -0.47, Welch t= -2.51 and Welch prob. =0.01) show DIF in favor of Sciences and Engineering groups and item 20 (DIF contrast= 0.44, Welch t= 2.28 and Welch prob. = 0.02) shows DIF in favor of Humanities group. The value of item 8 in DIF contrast column is less than 0.43; therefore, it has been categorized as having small (A) level of DIF. In items 20

and 23, the value of DIF contrast is more than 0.43 and less than 0.64, therefore they show moderate (B) DIF.

4.1.5. Vocabulary Skill Using Winsteps Program

As Table 6 demonstrates, 3 items of vocabulary skill have been flagged as DIF. They are item 39 (DIF contrast= 0.36, Welch t= 1.91 and Welch prob. = 0.05), item 41 (DIF contrast= 0.39, Welch t= 1.93 and Welch prob. = 0.05), and item 58 (DIF contrast= 0.61, Welch t= 2.44 and Welch prob. =0.01).

All DIF items have positive values in DIF contrast column and show DIF toward focal group (Humanities students). Items 39 and 41 have a value less than 0.43 in DIF contrast column and they have been categorized as having a small (A) level of DIF but the value of DIF contrast for item 58 is more than 0.43 and less than 0.64, so it has been categorized as having a moderate (B) level of DIF.

Table 5

DIF Flagged Items of Grammar Skill Detected by IRT Model

Name	PC1 DIF Measure	PC2 DIF Measure	DIF Contrast	Welch t	Welch prob.
Q8	-1.41	-1.04	-0.38	-2.04	0.0421
Q20	0.46	0.02	0.44	2.28	0.0233
Q23	-0.35	0.12	-0.47	-2.51	0.0126

Table 6

DIF Flagged Items of Vocabulary Skill Detected by IRT Model

Name	PC1 DIF Measure	PC2 DIF Measure	DIF Contrast	Welch t	Welch prob.
Q39	0.23	-0.13	0.36	1.91	0.0565
Q41	0.68	0.29	0.39	1.93	0.0547
Q58	1.54	0.93	0.61	2.44	0.0152

4.1.6. Reading Comprehension Skill Using Winsteps Program

As shown in Table 7, two items of reading skill display DIF. Item 64 (DIF contrast= 0.48, Welch t= 2.55 and Welch prob. = 0.01) and item 90 (DIF contrast= 0.58, Welch t= 3.21 and Welch prob. = 0.00) have positive values in DIF contrast column and show DIF toward the Humanities group. Both of them show greater values than 0.43 and fewer values than 0.64 in DIF contrast column. Therefore, they have been categorized as having a moderate (B) level of DIF.

Referring back to Tables 5, 6, and 7 and considering that three items of grammar skill (items 8, 20, and 23), three items of vocabulary skill (items 39, 41, and 58) and two items of reading skill (items 64 and 90) have been flagged as DIF, it is noticed that the number of DIF detected items in grammar and vocabulary sections is more than reading section; but because out of 30 items, only 3 items showed DIF in these sections, it is concluded that items of EPT do not show DSF in favor of a particular academic discipline group using IRT procedure.

Table 7

DIF Flagged Items of Reading Skill Detected by IRT Model

Name	PC1 DIF Measure	PC2 DIF Measure	DIF Contrast	Welch t	Welch prob.
Q64	-0.96	-1.44	0.48	2.55	0.0112
Q90	-0.15	-0.73	0.58	3.21	0.0015

4.2. Results of Question Two

To investigate DIF in this study across academic disciplines, the data was analyzed by DIFAS program. With this aim, the participants of Sciences and Engineering disciplines were taken as the reference group and the participants of Humanities discipline were taken as the focal group. As shown in Table 8, of all 100 items, ten items show DIF.

DIF detected items include item 23 (MH CHI= 5.95, BD= 0.00), item 26 (MH CHI= 0.09, BD= 5.12), item 38 (MH CHI= 0.33, BD= 9.95), item 41 (MH CHI= 6.46, BD= 0.98),

item 49 (MH CHI= 0.95, BD= 6.13), item 64 (MH CHI= 8.20, BD= 4.64), item 68 (MH CHI= 1.54, BD= 6.47), item 72 (MH CHI= 0.24, BD= 7.73), item 86 (MH CHI= 0.55, BD= 5.84), and item 90 (MH CHI= 11.51, BD= 0). The value of MH CHI for items 23, 41, 64, and 90 is more than 3.84 and the value of BD for items 26, 38, 49, 64, 68, and 86 is more than 3.84.

These items are from different skills. Items 38, 41, 64, 86, and 90 in MH LOR column have negative values, so they demonstrate DIF in favor of the focal group and items 23, 26, 49, 68, and 72 have positive values and manifest DIF in favor of the reference group. Items 23, 41, 64, and 90 have been categorized as having a moderate (B) level of DIF and items 26, 38, 49, 68, 72 and 86 have been categorized as having a small (A) level of DIF. In order to prove the existence of DIF items in the MH CHI column, LOR Z column is also studied and it shows 4 items with DIF, one of them (item 23) is more than +2 and three of them (items 41, 64, and 90) are less than -2. As shown in Table 8, five positive and five negative values in MH LOR column indicated a good balance between the reference and focal groups directed items.

Table 8

DIF Detected Items of EPT Using Mantel-Haenszel Procedure

NAME	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Q23	5.9533	0.4905	0.1952	2.5128	0.001	Flag	B
Q26	0.0969	0.0751	0.1865	0.4027	5.126	Flag	A
Q38	0.3394	-0.1597	0.2337	-0.6834	9.957	Flag	A
Q41	6.4622	-0.5441	0.2082	-2.6134	0.983	Flag	B
Q49	0.9555	0.1962	0.1855	1.0577	6.136	Flag	A
Q64	8.2091	-0.5598	0.1927	-2.905	4.646	Flag	B
Q68	1.5466	0.2574	0.1946	1.3227	6.471	Flag	A
Q72	0.2468	0.1394	0.2308	0.604	7.732	Flag	A
Q86	0.5549	-0.2146	0.2465	-0.8706	5.841	Flag	A
Q90	11.5106	-0.6513	0.1859	-3.5035	0	Flag	B

Table 9 demonstrates the results of IRT analysis for different academic discipline groups and their performance on different items of EPT by Winsteps program. For simplicity, only items displaying DIF are presented.

As shown in Table 9, of all 100 items, 8 items show DIF. DIF detected items include items 8, 20, 23, 39, 41, 58, 64, and 90. These items are from different sections of EPT. Of all these items, items 8 and 23 have negative values in DIF contrast column and they show DIF in favor of Sciences and Engineering groups. Items 20, 39, 41, 58, 64, and 90 have positive value, so they manifest DIF in favor of Humanities group. Among these items, the value of items 8, 39, and 41 in DIF contrast column is less than 0.43; therefore, they show small (A) DIF. The value of DIF contrast for other items is more than 0.43 and less than 0.64; so according to Table 4.4, they show moderate (B) DIF.

Table 9

DIF Detected Items of EPT Using IRT Procedure

Name	PC1 DIF Measure	PC2 DIF Measure	DIF Contrast	Welch t	Welch prob.
Q8	-1.41	-1.04	-0.38	-2.04	0.0421
Q20	0.46	0.02	0.44	2.28	0.0233
Q23	-0.35	0.12	-0.47	-2.51	0.0126
Q39	0.23	-0.13	0.36	1.91	0.0565
Q41	0.68	0.29	0.39	1.93	0.0547
Q58	1.54	0.93	0.61	2.44	0.0152
Q64	-0.96	-1.44	0.48	2.55	0.0112
Q90	-0.15	-0.73	0.58	3.21	0.0015

4.3. Results of Question Three

Four items flagged as DIF by IRT, have been flagged by Mantel-Haenszel method using DIFAS program. They are items 23, 41, 64, and 90. Items 64 and 90 show moderate (B) DIF in favor of Humanities group, item 23 shows moderate (B) DIF and it advantaged the

reference group (Sciences and Engineering) by MH and IRT analyses using DIFAS and Winsteps, respectively; and finally, item 41 has been categorized as having small (A) DIF in favor of Humanities group using IRT procedure and having moderate (B) DIF in favor of Humanities group using MH procedure. Items 26, 38, 49, 68, 72, and 86 are DIF items detected only by MH method using DIFAS program. Among these items, items 49, 68, and 72 show small (A) DIF in favor of Sciences and Engineering groups; items 38 and 86, show small DIF in favor of Humanities group and item 26 has been categorized as having moderate (B) DIF in favor of Sciences and Engineering groups. Items 8, 20, 39, and 58 are items flagged as DIF by IRT procedure using Winsteps program.

Table 10

DIF Items Detected by Mantel-Haenszel and IRT

Item	DIF Size by MH/advantaged group		DIF Size by IRT/advantaged group	
Q8	-	-	A	Sciences & engineering
Q20	-	-	B	Humanities
Q23	B	Sciences & engineering	B	Sciences & engineering
Q26	B	Sciences & engineering	-	-
Q38	A	Humanities	-	-
Q39	-	-	A	Humanities
Q41	B	Humanities	A	Humanities
Q49	A	Sciences & engineering	-	-
Q58	-	-	B	Humanities
Q64	B	Humanities	B	Humanities
Q68	A	Sciences & engineering	-	-
Q72	A	Sciences & engineering	-	-
Q86	A	Humanities	-	-
Q90	B	Humanities	B	Humanities

Item 8 shows small (A) DIF in favor of Sciences and Engineering groups, items 20 and 58 have been categorized as having B DIF in favor of Humanities group and item 39 shows A DIF and advantaged the Humanities group. Table 10 illustrates the DIF size of items

and the group has been advantaged by each item; it also shows whether flagged items as DIF have been detected by both methods or not. In order to see whether there is any correlation between these two methods, SPSS is used and as shown in Table 10, $p<0.05$ and it is proved that there is a positive correlation between MH and IRT methods.

Table 11

Correlation between IRT and MH Methods

		IRT	MH
IRT	Pearson Correlation	1	-.548*
	Sig. (2-tailed)		.043
	N	14	14
MH	Pearson Correlation	-.548*	1
	Sig. (2-tailed)	.043	
	N	14	14

**correlation is significant at the 0.05 level (2-tailed)*

4.4. Results of Question Four

In order to investigate whether DIF items are biased or not, the researchers used content analysis. Item 8 belongs to part one of grammar section. It is an incomplete sentence and students have to choose the one word that best completes the sentence.

Item 8. *Usually strokes happen ……..a combination of factors, such as certain medical conditions, inherited characteristics or unhealthy sounds.*

A. for
B. of
C. because of
D. because

The value of Welch probability for this item is 0.04 using IRT method. It is less than 0.05 and flagged as DIF. It is considered as having a small level of DIF because the value of DIF Contrast is -0.38 and according to Table 4, it is less than 0.43. Its negative value also indicates that the item favors Sciences and Engineering groups. Content analysis showed that this item has been related to Sciences groups and because of this, the Humanities group found it harder to choose the best word. Therefore, it can be concluded that DIF item constitutes bias and items function differently for two groups of test takers because they are biased.

Items 20, 23, and 26 belong to the grammar section related to written expressions. The participants have to identify the one underlined word phrase that must be changed in order for the sentence to be correct.

Item 20. *The information on the Iranian panther is too limited since this animal is regarded*
A B C D
as very shy in the wildlife.

This item detected as DIF, only by IRT method. The value of DIF contrast for this item is 0.44 and it is more than 0.43 and less than 0.64, so it demonstrates moderate DIF and positive value in DIF contrast column shows that this item is in favor of Humanities group, although the content of this item is more related to Biology courses. So, this DIF item does not demonstrate bias and it functions differently for two groups of test takers simply because they differ in their ability.

Item 23. *Stars in our universe vary in temperature, color, bright, size, and mass.*
A B C D

This item demonstrates DIF using MH and IRT methods by DIFAS and Winsteps program; but using DIFAS program, the value of MH CHI is 5.95 and it is more than 3.84, and ETS, flagged it as moderate DIF. The positive value in MH LOR column shows that it is in favor of Sciences and Engineering groups. Using IRT, it shows moderate DIF in favor of Sciences and Engineering groups (Welch prob. = 0.01 and DIF contrast= -0.47). Content analysis showed that this item is more related to Physics and Geology courses belonging to Sciences academic discipline and because it detected as DIF using both methods, it can be concluded that item 23 displays bias.

Item26. *Dependence on the text messaging has spread rapid since mobile phones invaded*
A B C D.
the market at the end of the last century.

The value of BD for this item is estimated as 5.12 by MH method using DIFAS software and it displays moderate amount of DIF. The positive value for MH LOR column indicates that this item has DIF in favor of Sciences and Engineering groups. Content analysis showed that this item is free from bias because it depends on no academic discipline and it functions differently, simply because two groups of test takers differ in their ability.

Items 38, 39 and 41 are related to part one of vocabulary skills. In these items, some words have been underlined. Four choices, marked (A), (B), (C), and (D), have been suggested as their synonyms and participants have to choose one of these choices that is the most appropriate synonym for the underlined word.

Item 38. *Presumably desolate desert and tundra areas actually harbor many forms of life.*

A. favor

B. halt

C. generate

D. shelter

This item flagged as DIF only by MH estimated by DIFAS program. The value of BD for this item is 9.957 and it is more than 3.84. According to ETS, it has been categorized as small DIF. The value of MH LOR is -0.15 and it shows that this item is in favor of Humanities group. It has been shown in content analysis that this item is more related to Geography courses belonging to Humanities academic discipline, so it constitutes bias.

Item 39. *Due to large deposits of debris left by glaciers, the Mid western plain of United States is an extremely fertile area for farming.*

A. level

B. rich

C. spacious

D. arid

The value of Welch prob. for this item is 0.05 using IRT method, so it is flagged as a DIF item. According to Table 4, the value of 0.36 in DIF Contrast column indicates that this item has been categorized as having small DIF, because it is less than 0.43. The positive value of DIF Contrast for this item shows that it is in favor of Humanities group. It means that this item is hard for Sciences and Engineering groups. Content analysis showed that the content of this item is related to Geography and Humanities academic discipline, therefore, it can be concluded that item 39 constitutes bias and because of this, it was easier for the Humanities group.

Item 41. *Poorly constructed dwellings cannot withstand severe storms.*

A. Houses

B. Piers

C. Boats

D. roads

This item demonstrates DIF using MH and IRT method by DIFAS and Winsteps program, but the value of MH CHI is 6.46 using DIFAS program and it is more than 3.84, and ETS, flagged it as moderate DIF. The negative value in MH LOR column shows that it is in favor of Humanities group. Using IRT, it displays small DIF in favor of Humanities group (Welch prob. = 0.05 and DIF contrast= 0.39); however, content analysis showed that it is related to Engineering academic discipline. Therefore, it can be concluded that this item does not display bias and it functions differently just because the two groups differ in their ability.

Items 49 and 58 belong to part two of the vocabulary section, one word is removed from questions and four choices have been suggested for it in these items. The participants have to choose the best option.

Item 49. *A law that _____ tobacco advertising in newspapers and magazines has just been made public.*

A. prohibits

B. recycles

C. surrenders

D. overcomes

This item was detected as DIF only by MH method using DIFAS program. The amount of BD for this item is 6.13 and it is more than 3.84. The ETS column shows that this item has been categorized as having a small level of DIF. The value of MH LOR for this item is 0.19 and it indicates that the Sciences and Engineering group is advantaged; although the content analysis indicated that this item is more related to Law courses and Humanities academic discipline, it can be said that this DIF detected item does not display bias.

Item 58. *We can_____ the power of the wind to generate electricity.*

A. harness

B. justify

C. engender

D. obey

The value of Welch prob. for this item is 0.01 using IRT method. The value of DIF contrast for this item is 0.61 and it displays a moderate amount of DIF. Its positive value in DIF contrast columns indicates that the item favors Humanities group. It means that Sciences and Engineering group find it harder to choose the best option; however, the content of this item is more related to Engineering academic discipline. So it can be concluded that this item does not show bias.

Items 64, 68, 72, 86, and 90 belong to the reading comprehension section. These items are related to different passages. In these items, the participants have to read passages and answer the following questions. Item 64 belongs to passage one.

Item 64. *The word "vogue" in line 14 is closest in meaning to……..*

A. tendency

B. prevalence

C. challenge

D. marvel

Using DIFAS and Winsteps program, this item demonstrates DIF for both methods. But the value of MH CHI is 6.46 and it is more than 3.84 using DIFAS program. It is flagged as moderate DIF. The negative value in MH LOR column shows that it is in favor of Humanities group. Using IRT procedure with Winsteps program, it shows moderate DIF in favor of Humanities group (Welch prob. = 0.01 and DIF contrast= 0.48). Content analysis showed that the content of the first passage is about Arabic language and it is related to Humanities academic discipline and because of this, item 64 is easier for Humanities students rather than Sciences and Engineering students. Therefore, it can be concluded that this item constitutes bias. Items 68 and 72 are related to passage two.

Item 68*. What is the main point of the first paragraph?*

A. The waves created by ocean currents are very large.

B. Despite the strength of the wind, it only moves surface water.

C. Deep ocean water is seldom affected by forces that move water.

D. The tides are the most powerful force to affect the movement of ocean water.

The value of BD for this item is 6.47 using MH method by DIFAS and it is more than 3.84. This item is flagged as small DIF. The positive value of MH LOR for this item indicates that it is toward Sciences and Engineering group. It means that this item is hard for the Humanities group. Content analysis indicated that the passage to which this item belongs is more related to Physics and Geology belonging to Sciences academic discipline, so this item constitutes bias.

Item 72. *The word "configuration" in line 19 is closest in meaning to……..*

A. unit

B. center

C. surface

D. arrangement

The value of BD for this item is 7.73 using MH method by DIFAS and it is more than 3.84. This item is flagged as small DIF. The positive value of MH LOR for this item indicates that Sciences and Engineering group is advantaged. Analyzing the content of this item showed that the text to which this item belongs is more related to Physics and Geology belonging to Sciences academic discipline, so this item constitutes bias. Item 86 belongs to passage four.

Item 86. *The phrase "occupy the spotlight" in line 23 is closest in meaning to*

A. receive the most attention

B. go the furthest

C. conquer territory

D. lighten the load

Using MH method done by DIFAS program, the value of BD for this item is estimated as 5.84 and because it is more than 3.84, this item is flagged as DIF. The value of MH LOR for this item is -0.21 and according to ETS column, it has been categorized as having a small level of DIF. Content analysis indicated that this item must be in favor of the Sciences group because it is more related to Sciences courses but its negative value in MH LOR showed DIF in favor of the Humanities group. Therefore, it can be concluded that this item does not constitute bias. Item 90 belongs to passage five.

Item 90. *What aspect of drama does the author discuss in the first paragraph?*

A. The reason drama is often unpredictable

B. The seasons in which dramas were performed

C. The connection between myths and dramatic plots

D. The importance of costumes in early drama

This item has been flagged as DIF using MH and IRT methods by DIFAS and Winsteps software. Using DIFAS program, the value of MH CHI is 11.51 and it is more than 3.84, and ETS, flagged it as moderate DIF. The negative value in MH LOR column shows that it is in favor of the Humanities group. Using IRT and MH procedures with Winsteps program, we can show moderate DIF which indicates that Humanities group is advantaged (Welch prob. = 0.00 and DIF contrast= 0.58). Content analysis indicated that the passage to which this item belongs is more related to Literature belonging to Humanities, so this item displays bias.

5. Discussion

In this study, we examined the EPT for DIF attributable to academic disciplines. A two-step approach was used: First, the test items were examined for DIF, and second, the items that were flagged were subjected to content analysis by the researchers. In following sections, the overall findings of DIF/ DSF analysis are discussed.

Out of 100 items, 10 items were flagged as DIF in the statistical analysis. Among these, 2 items belonged to the grammar section, 3 items belonged to the vocabulary section, and 5 items belonged to the reading sections. The number of DIF items in the reading part is

more than the two other parts, but because only 5 items out of 40 were flagged as DIF (12.5%) and out of these 5 items, three of them indicated that the Humanities group is advantaged and two of them indicated that the Sciences and Engineering group is advantaged, it is concluded that items of EPT do not show DSF in favor of a particular academic discipline by using MH procedure.

As shown in Tables 5, 6, and 7, three items of grammar skill (10%), three items of vocabulary skill (10%), and two items of reading skill (5%) have been flagged as DIF. Among these three DIF detected items belonging to grammar skill, 2 items advantaged Sciences and Engineering group and one item advantaged Humanities group. Hence, 10% of grammar items show DIF and these items were not in favor of a particular academic discipline and cancel out each other, it can be concluded that grammar skill does not show DSF. Although all DIF detected items of vocabulary skill advantaged Humanities group, 3 items out of 30 are too small to consider this skill as having DSF. Several researchers examined the potential effects of DIF and DSF cancellation at test level in this regard (See e.g., Drasgow, 1987; Maller, 2001; Nandakumar, 1993; Roznowski, 1987; Roznowski & Reith, 1999). Investigating research question one, it may suggest that there existed some cancellation effects at test level because the expected score differences for the Humanities and Sciences were rather small in comparison to the number of items identified as indicating DIF for or against each academic category (Drasgow, 1987).

Investigating research question two, ten out of 100 items of EPT (10 percent) exhibited DIF by academic discipline. Out of these, fifty percent showed DIF in favor of the Humanities group and fifty percent showed DIF in favor of the Sciences and Humanities groups. Five items exhibited moderate (B) DIF and the others showed small (A) DIF. Therefore, it is concluded that the items of EPT show DIF in favor of a particular academic discipline group using Mantel-Haenszel procedure. However, previous research suggests that approximately one-third to one-half of personality test items show potential DIF. Waller et al. (2000) discovered potential DIF by race (Black vs. Caucasian) on 38% of the MMPI questions, while Sheppard et al. (2006) discovered potential DIF by race/culture (Chinese vs.

American) on 58% of the items in the Revised NEO Personality Inventory's conscientiousness scale (NEO-PI-R).

IRT identified 8 out of 100 items with DIF. Among these, twenty five percent indicated DIF in favor of the Sciences and Engineering group and seventy five percent advantaged the Humanities group. Five of these DIF detected items (62.5%) exhibited moderate (B) DIF and three of them (37.5%) exhibited small (A) DIF. As a result, it can be concluded that the EPT which was the object of this investigation shows DIF toward a particular group included in this study using IRT procedure. Since no studies to date have identified DIF in the EPT, the findings of this study have limitations in comparison across studies and require further investigations. For example, reviewing these items by gender and academic discipline as well as querying the students qualitatively may assist in understanding the differences between the group performances; However, as the first DIF study on the EPT, it is important to describe for future explanation that there were fourteen DIF items in the EPT across academic disciplines in the context of universities in Tehran. Thus, the revision or elimination of DIF items should be done with caution because they can represent important content for the future test construction. Using SPSS, it was found that there was a positive correlation between MH and IRT procedures.

For test development, DIF analysis is a useful tool. When significant amounts of DIF are found and eliminated from tests, they are more likely to be valid and less biased. According to the results of the content analysis across the two subscales, items dealing with science-related issues were differentially easier for the Sciences, whereas items addressing human interactions were differentially easier for the Humanities. Items 8 and 39 showed small (A) DIF using IRT method. Item 8 advantaged Sciences and Engineering group and item 39 advantaged Humanities group. They are also detected as DIF items by content analysis. Items 38, 68, and 72 were detected as small (A) DIF using MH procedure. The first one was in favor of Humanities group and the last two items belonging to passage two of the reading section, indicated DIF in favor of the Sciences and Engineering group. The content of the text is about tides and waves in the sea and the ocean which is more related to Physics and Geology belonging to Sciences academic discipline, so this item shows bias. Items 23,

64, and 90 exhibited moderate (B) DIF using IRT and MH procedures. Item 23 advantaged Sciences and Engineering group and the last two items belonging to passage one and passage five of the reading section, advantaged Humanities group. The content of the passage one is about Arabic language and the content of passage five is about drama. Hence, both of them are related to Humanities courses, it is concluded that these items display bias and it is better to modify them. This pattern contrasts with findings from gender DIF research, which found that females performed better than a matched group of males on reading items pertaining to human connections, but males were more likely to outperform a comparable group of females on science-related issues (e.g., Carlton & Harris, 1992; Curley & Schmitt, 1993; Lawrence & Curley, 1989; Lawrence et al., 1988; Maller, 2001; O'Neill & McPeek, 1993; O'Neill et al., 1993; Scheuneman & Gerritz, 1990).

However, some items were not consistent with the DIF pattern, which may limit the generalization of the findings. Items 20 and 58 indicated moderate (B) DIF in favor of the Humanities group using IRT method, although content analysis showed that these items are more related to Sciences courses and it is better to say that they are in favor of the Sciences group. Items 49 and 86 detected as small (A) DIF using MH procedure. The result of DIF procedures is not consistent with the result of content analysis. The content of item 49 is about law belonging to the Humanities group, but using MH, this item advantaged the Sciences group. Item 86 belongs to passage four which is related to diseases and medicine, so it is better to say that this item is in favor of the Sciences group but using MH, it has been shown that this item advantaged the Humanities group. Item 41 was detected as moderate (B) DIF and small (A) DIF using MH and IRT, respectively. Using DIF procedures, this item advantaged the Humanities group, although content analysis showed that the text of this item is related to Engineering academic discipline. Content analysis of item 26 indicated that this item is not in favor of a particular academic discipline but it was detected as moderate (B) DIF in favor of Sciences and Engineering group by using MH procedure. As numerous studies have argued, DIF items that deviate from the conventional DIF pattern demonstrate that content analysis alone is not a reliable means of detecting DIF (e.g., Engelhard et al., 1990; Plake, 1980; Rengel, 1986; Sandoval & Miille, 1980). However, a fair evaluation is necessary to make judgmental analyses of items that stereotype and under-represent a

particular population of examinees, because those items may cause the subgroup to be less motivated to do well on the test (Clauser & Mazor, 1998; Tittle, 1982).

In a similar vein, Scheuneman and Gerritz (1990) found that item content alone is not a reliable predictor of DIF direction, implying the inclusion of other unknown elements that lead to DIF. A large DIF value indicates the presence of additional structures that may function differently across the reference and focal groups because DIF approaches are based on comparable examinees.

6. Conclusion

This study investigated academic discipline differences in the English Proficiency Test (EPT) among 642 post-graduate students who were going to pursue their Ph.D. studies in a variety of academic disciplines including Humanities, Sciences, and Engineering. The results revealed that there were fourteen DIF items (4 grammar items, 5 vocabulary items, and 5 reading comprehension items) in the EPT across academic disciplines with Humanities group reporting a higher grade than Sciences and Engineering groups using Mantel-Haenszel (MH) and Item Response Theory (IRT) procedures. Four items (items 23, 41, 64, and 90) have been flagged as DIF by MH and IRT procedures, 6 items detected as DIF using only MH procedure (items 26, 38, 49, 68, 72, and 86) and 4 items have been flagged as DIF items only by IRT method (items 8, 20, 39, and 58). Six items have been categorized as having moderate (B) DIF (items 20, 23, 26, 58, 64, and 90), seven items have been categorized as having small (A) DIF (items 8, 38, 39, 49, 68, 72, and 86) and one of them which is flagged as DIF by both procedures, showed small DIF by IRT and moderate DIF by MH method (item 41). Winsteps and DIFAS software were utilized in this study to detect DIF using IRT and MH procedures, respectively. This study showed that the items of EPT do not show DSF in favor of a particular academic discipline using both MH and IRT methods.

SPSS was used to determine Mantel-Haenszel and Item response theory procedures correlate with each other and the result showed that there is a positive correlation between IRT and MH.

To see whether DIF detected items are biased or not, the researchers did content analysis. Interpreting DIF items in educational assessment is not an easy task, and as our attempt to find an explanation in the content analysis illustrates, it is even more difficult in the context of second language testing. Nonetheless, because test content and the types of tasks posed by particular test items are invariably major determinants of test performance, we believe that a closer analysis of the relationships between item content and DIF is one way for conducting future research. The result of the content analysis showed that out of 14 items, 8 items (items 8, 23, 38, 39, 64, 68, 72, and 90) are biased and 3 of these items were flagged as DIF using MH and IRT methods. Content analysis showed that 6 items are not biased; the content of item 26 is not related to a particular academic discipline, so it is free from bias, but the content analysis of 5 other items showed a different advantaged group from the one determined by the statistical procedures. So it cannot be explained precisely whether these items constitute bias or not, but it can be concluded that they are not biased and they function differently because the two groups differ in their abilities.

To summarize, the goal of a DIF analysis is to guarantee that test equity or fairness is met. The statistical identification of items that show signs of DIF is a critical step in achieving this goal. Because tests are inherently multidimensional, and multidimensionality is the main cause of DIF, a better understanding of test dimensionality and the effects of these dimensions on DIF could lead to a more accurate test score interpretation, greater control over the influence of relevant auxiliary dimensions, and a reduction in the influence of unintended and irrelevant nuisance dimensions. Using the DIF detection approach, this study has made a first contribution to the issue of test fairness for all test takers, regardless of their academic subject. Although it needs to be replicated with different samples, preliminary content analysis of those items flagged for DIF reveals that DIF may be connected with content characteristics particular to group membership (e.g., science-related topics or human issues). It should be emphasized that the preliminary content analysis described in this study was based on the researchers' subjective opinion rather than that of a panel of content experts, and thus could be subject to misclassification. To make the measurement procedure as precise as feasible, this study identifies the need for regular evaluation and revision of tests. The researchers' main recommendation is for test developers and designers. Test

developers play critical roles in the assessment process. They should be aware of the negative effects of DIF and DSF on test validity and test results so as to design and construct tests which do not advantage any groups by considering as many factors as possible. The results of this study offer test developers and designers to develop tests free from bias and pay attention to choice of test topics so that as far as possible no academic discipline group is clearly disadvantaged. It is also recommended that the items displaying significant DIF be analyzed and revised.

Conflict of interest

The author(s) certify/certifies that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in the present research paper.

References

Abbott, M. L. (2016). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24*(1), 7-36. DOI: 10.1177/0265532207071510.

Abedi, J., Bailey, A.L., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2005). *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation from Three Perspectives* (CSE Report 663). Retrieved from http://cse.ucla.edu/products/reports/r663.pdf#page=87.

Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly, 5*(2), 120-141. DOI: https://doi.org/10.1177/026553228500200204.

Azocar, F., Arean, P., Miranda, J., & Muñoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology, 57*(3), 355-365. DOI: 10.1002/jclp.1017.

Bachman, L. F. (2004). *Statistical analyses for language testing*: Cambridge, UK: Cambridge University Press.

Camilli, G. (2006). Test fairness. *Educational Measurement, 4*, 221-256. DOI: https://doi.org/10.1002/jclp.1017.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: *Sage*.

Carlton, S.T., & Harris, A.M. (1992). Characteristics associated with Differential Item Functioning on the Scholastic Aptitude Test: gender and majority /minority group comparisons (ETS Research Report, 92–64). Princeton, NJ: ETS Research Report. DOI: doi/pdf/10.1002/j.2333-8504.1992.tb01495.x

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155-163. DOI: https://doi.org/10.1177/026553228500200204

Clapham, C. (1996). *The development of IELTS: A study of the effect of background on reading comprehension.* Cambridge: Cambridge University Press.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44. DOI: https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Curley,W.E., &Schmitt, A.P. (1993). Revising SAT-Verbal items to eliminate Differential Item Functioning. New York: College Entrance Examination Board.

De Ayala, R. J., Kim, S., Stapleton, L. M., & Dayton, C. M. (1999, July). *A reconceptualization of differential item functioning.* Paper Presented at the Annual Meeting of American Educational Research Association (AERA), Montreal, Canada.

Dodeen, H., & Johanson, G. A. (2003). An Analysis of Sex-related Differential Item Functioning in Attitude Assessment. *Assessment & Evaluation in Higher Education, 28*(2), 129-134. DOI: 10.1080/02602930301667.

Dorans, N., & Holland, P.W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P.W. Holland, & H. Wainer (Eds), *Differential Item Functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum. DOI: 10.21031/epod.368081.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*(1), 19.

Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and Modern Greek. *Language Learning, 46*(2), 233-282.

Engelhard Jr, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*(4), 347-360.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination. *Language Assessment Quarterly, 4*(2), 190-222. DOI: 10.1080/15434300701375758.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46*(3), 314-329.

Hale, G. A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing, 5*(1), 49-61.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

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

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16*(3), 385-402. DOI:10.1007/BF03173189.

Koo, J., Becker, B. J., & Kim, Y. S. (2013). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing, 31*(1), 89-109. DOI: 10.1177/0265532213496097.

Lawrence, I.M., & Curley, W.E. (1989). Differential Item Functioning for males and females on SAT-Verbal Reading subscore items: follow-up study (ETS Research Report 89–22). Princeton, NJ: Educational Testing Service.

Lawrence, I.M., Curley, W.E., & McHale, F.J. (1988). Differential item functioning for males and females on SAT verbal reading subscore items (Report No. 88–4). New York: College Entrance Examination Board.

Li, H., & Suen, H. K. (2012). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing, 30*(2), 273-298.

Linacre, J. (2010). *Winsteps Rasch measurement: Computer software manual (version 3.70. 0.2)*. Retrieved from http://www.winsteps.com

Liu, I. M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, *52*, 1223-1234.

Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement, 46*(9), 3228-3237. DOI: 10.1016/j.measurement.2013.06.020.

Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology, 35*(3), 299-314.

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*(5), 793-817.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719-748.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451.

McNamara, T. F. (1996). *Measuring second language performance*. London: Addison Wesley Longman.

McNamara, T. F.,, & Roever, C. (2006). *Language testing: The social dimension*. Madison: John Wiley & Sons.

Nandakumar, R. (1993). Simultaneous DIF Amplification and Cancellation: Shealy-Stout's Test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). New Jersey: Lawrence Erlbaum Associates, Inc.

O'Neill, K.A., McPeek, W.M., & Wild, C.L. (1993). *Differential Item Functioning on the Graduate Management Admission Test* (ETS Research Report 93–35). Princeton, NJ: ETS.

Pae, T. I. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21*(1), 53-73.

Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing, 29*(4), 533-554.

Paek, I., & Wilson, M. (2011). Formulating the rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023-1046.

Park, G. P., & French, B. F. (2013). Gender differences in the Foreign Language Classroom Anxiety Scale. *System, 41*(2), 462-471. DOI: 10.1016/j.system.2013.04.001.

Penfield, R. D. (2001). Assessing Differential Item Functioning Among Multiple Groups: A Comparison of Three Mantel-Haenszel Procedures. *Applied Measurement in Education, 14*(3), 235-259. DOI: 10.1207/s15324818ame1403_3.

Penfield, R. D. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *Alberta Journal of Educational Research, 49*(3), 231-243.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti Estimator of the Cumulative Common Odds Ratio to DIF Detection in Polytomous Items. *Journal of Educational Measurement, 40*(4), 353-370. DOI: https://doi.org/10.1207/S15324818AME1403_3.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40*(2), 397-404.

Rengel, E. (1986). Agreement between statistical and judgmental item bias methods. Retrieved from ERIC database. (ED289 890)

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.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72*(3), 480-483.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*(2), 248-269.

Runnels, J. (2013). Measuring differential item and test functioning across academic disciplines. *Language Testing in Asia, 3*(1), 12-29.

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

Sandoval, J., & Miille, M. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48*(2), 249-253.

Saville, N. (2005). Setting and monitoring professional standards: A QMS approach. *Cambridge ESOL Research Notes,* 22, 2-5.

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131.

Schmitt, A. P., & Dorans, N. J. (1990). Differential Item Functioning for Minority Examinees on the SAT. *Journal of Educational Measurement, 27*(1), 67-81.

Schmitt, N., Hattrup, K., & Landis, R. S. (1993). Item bias indices based on total test score and job performance estimates of ability. *Personnel Psychology, 46*(3), 593-611.

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.

Shohamy, E. (2000). Fairness in language testing. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 15–19). Cambridge: Cambridge University Press.

Thissen, D., Steinberg, L., &Wainer, H. (1988). Use of item response theory in the study of group differencesin trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In R.A. Berk (Ed.), *Handbook of methods of detecting test bias* (pp. 31–30). Baltimore, MD: Johns Hopkins University Press.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing, 22*(2), 211-234.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: an illustration with the MMPI. *Psychological Methods, 5*(1), 125-146.

Yu, L., Lei, P. W., & Suen, H. K. (2006, April). *Using a differential item functioning (DIF) procedure to detect differences in opportunity to learn (OTL).* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa, Canada: National Defense Headquarters.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement (Research Report ETS RR-12-08). Princeton, NJ: Educational Testing Service.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321-344.