Research in English Language Pedagogy (2022)10(1): 169-190

DOI: 10.30486/RELP.2021.1935588.1295

Original Article

Examining Differential Item Functioning (DIF) For Iranian EFL Test Takers with Different Fields of Study

Shokouh Rashvand Semiyari[1,*], *Saeideh Ahangari*[2]

[1]Department of English Language Teaching, East Tehran Branch, Islamic Azad University, Tehran, Iran

[2]Department of English Language Teaching, Tabriz Branch, Islamic Azad University, Tabriz, Iran

Submission date: 14 Jul, 2021 Acceptance date: 13 Oct, 2021

Abstract

Differential Item Functioning (DIF) takes place when different groups of test-takers with the same level of ability perform differently on a single test. It means some other factors might arise due to group membership. The object of this article was to examine DIF in the MSRT (MCHE) test items. This is an English proficiency test that comprises a total of 100 questions including listening comprehension (LC), structure and written expressions (SWE), and reading comprehension (RC) sections. To this end, 200 pre-intermediate to intermediate Iranian EFL learners with the age range of 25 to 32 in two different fields of study (100 Humanities and 100 sciences) were randomly selected for the analysis. The Item Response Theory (IRT) Likelihood Ratio (LR) approach was used to identify items displaying DIF. The scored item of 200 test-takers was subjected to the IRT Three-Parameter Model presenting the probability that a randomly selected test taker with an ability of theta (θ) answered an item correctly, using item difficulty (b parameter), item discrimination (a parameter), and pseudo-guessing (c parameter). The results of the independent samples t-test for means comparison of two groups indicated that Science test-takers outperformed the Humanities, especially in SWE and RC sections. It can be inferred that the exam was statistically easier for the Science test-takers at 0.05 level. The findings identified 15 DIF items as well. The implications and suggestions for further studies were also reported.

Keywords: Differential Item Functioning (DIF), Fields of study, Item Response Theory (IRT), Likelihood ratio Approach (LR), MSRT (MCHE) Proficiency Test

Corresponding Author's E- mail: Sh.Rashvand@iauet.ac.ir / Sh_Rashvand@yahoo.com

1. Introduction

The development of psychometric tests and testing procedures has been influenced due to political and social changes within a few past decades (Owen, 1998). When it comes to psychometric tests and individual or group comparisons, item/test bias needs to be taken into account to minimize inappropriate interpretations. Test bias is different from test fairness in that it is usually assessed objectively while test fairness is some kind of assessment that is done subjectively and might not be explained in absolute concepts. It implies one cannot classify tests as either fair or not fair and it is only a matter of degree.

It is not the test characteristics which is significant by themselves but the test scores' interpretations and the outcomes that may arise are of great importance as the examinees' educational fates are usually determined by these decisions. The term 'biased' has to do with the instruments that are applied, testing procedures that are used, and the ways the tests are scored and interpreted. Bias does not merely refer to score differences between two groups (Osterlind, 1983). It has been substituted by differential item functioning (DIF) indicating that individuals who are similar due to their level of ability perform differently on a test and gain different scores accordingly. Test bias or DIF deals with systematic errors and reveals the features relating to item psychometric characteristics displaying the items cannot assess fairly considering different individuals/groups.

DIF arises when individuals from various classes, share the same level of ability but display different likelihood in responding to an item accurately. In essence, non-DIF depicts the situation in which test-takers with the same level of ability regardless of their in-group differences have the same chance to answer an item accurately. DIF refers to the extent to which the exam items discriminate between examinees with the same ability level from different groups including gender, ethnicity, education, etc. (Zumbo, 2007). Factors contributing to item/test bias are culture, education, language, socioeconomic status, and so on (Van de Vijver, 1998). Test bias or DIF needs to be checked during the test construction process. In fact, tests should be so meticulously constructed that when variability occurs in examinees' test results, such disparity could be attributed to differences in whatever the test is going to assess (Osterlind, 1983). By detecting and removing items indicating DIF, test makers will be able to develop more practical tests. Item analysis helps test makers detect problematic items that lack required psychometric properties. In this paper, item analysis was served by means of item response theory (IRT) based on the DIF study.

As confirmed by Li and Zumbo (2009), DIF can be observed in any kind of assessment-due to the fact it might not be either applied as part of the item analysis or can be easily ignored by researchers as a statistical decision method. Regardless of how DIF items take place, it is still unclear how these items affect the subsequent statistical outcomes and conclusions drawn. Moreover, a huge amount of research has been carried out on high-stakes tests to detect items showing DIF (eg, Hope et al., 2018; Oliveri et al, 2017). However, few studies have been conducted on the examinees' field of study in an Iranian context to date. To address such a gap and determine the impact of DIF items on the statistical conclusions from the examinees' test scores, the researchers carried out the study using IRT 3PL Model to detect items displaying DIF.

2. Literature Review

2.1. Methods for identifying DIF

DIF has to do with examinees' scores on the tests, their latent trait's (ability) evaluation, and investigation of individuals who are similar with regard to their level of ability (i.e., the individuals who come from various classes; yet perform similarly on an item). DIF analysis that stems from educational assessment has been widely applied in psychometric investigations to evaluate whether the likelihood of answering an item displays various statistical properties for different groups of examinees (Mousavi & Krishnan, 2016). As pointed out by Rezaee and Shabani (2010), using DIF identifying methodologies can assist to recognize the contributing factors to examinees' differential performance. Therefore, items displaying DIF can be removed and more accurate decisions would be made.

The methods which are taken for discerning DIF might vary according to the way the examinees are sort out. Three common approaches for detecting DIF are Mantel- Haenszel x^2 Test put forward by Mantel and Haenszel (1959). Such a test suits well even for a small number of participants. Moreover, it allows the test developers to use simple arithmetic procedures based upon logistic regression methods proposed by Zumbo (2007). Simple arithmetic procedures provide a more thorough description of DIF and thus would enable the researchers to make a distinction between uniform and non-uniform DIF. The other procedures considering IRT models have~~n~~ been stated by Lord, (1980), Raju (1990), and Thissen et al. (1994). These methods deal with examinees' level of ability and item characteristics and can be applied with~~a~~ larger sample size. Among these models, IRT was applied by the researchers to detect items flagging DIF, since IRT Model presents the most useful data for identifying differences on particular items (Ertuby, 1996).

2.2. Item Response Theory (IRT) Models

Most of the measurement procedures have focused on the latent variables (Hambleton, 1996). The chance of getting the right answer depends upon both item characteristics and examinees' level of ability. Such relation is mathematically expressed as item characteristic curve (ICC). Any ICC needs to predict the examinees' scores based on their underlying abilities. This is also known as the item response function. The examinees' level of abilities is shown along with the X-axis and it is shown by theta (θ) while the likelihood of responding to items accurately is represented on Y-axis and is shown by p (θ). Every item has its own ICC. As Baker (1985) proposed, the ICC shape relies upon the item difficulty (b-parameter), item discrimination (a-parameter), and guessing power known as pseudo-chance (c-parameter). In fact, ICCs might vary based on horizontal location displaying the individual level of ability and standing for item difficulty. The likelihood of choosing the right answer is 0.50 (i.e., the likelihood of choosing the right answer is 50 percent). Larger b-values stand for more difficult items. The b-value ranges from -2.5 to +2.5 in theory. That means it varies from very easy items to very tough ones. Item discrimination (a-parameter) shows the slope of the ICC and the precision of measurement of an item. The curve slope and item discrimination are positively correlated in the sense that the steeper slope shows more discriminating power of an item. The a-value ranges between 0~2. Those below 0.5 do not have discriminating power. The items having bigger discrimination power can well discriminate between the individuals. The guessing power (c-parameter) shows the likelihood a test taker with the lowest level of ability answered the item accurately. Items should have a multiple-choice format to make guessing possible. The c-parameter ranges from 0 to 1.

2.2.1. The Three General IRT Models: Basic features

IRT models change due to the properties of items they encompass. The one parameter or Rasch model has to do with item difficulty. The test items that do not fit the Rasch model are prone to revisions, deletion, and modifications. One of the vantages of the

Rasch model is that it makes a hypothetical unidimensional line along which items are maneuvered based on examinees' level of ability and item difficulty. The items that are close enough to the hypothetical line are related to the Rasch dimension and those items that fall far from the line are assessing another irrelevant dimension (Baghaei, 2008). Rasch models are somehow robust, however, the data are disorganized and imperfect and may never totally fit the model (Farrokhi & Esfandiari, 2011). The two-parameter model deals with item discrimination. Item difficulty plus item discrimination (probability of getting the correct response based on examinees' ability level) are taken into account. The third parameter or pseudo-chance parameter is realized when items have a multiple-choice format so that examinees can get the correct response by guessing. IRT models are unidimensional and independent. They are based upon the shape of ICC and examinees' level of ability.

2.3. Uniform vs. Non-uniform DIF

DIF falls within two distinct categories according to the logistic regression model: uniform and non-uniform. Uniform DIF influences the examinees at all levels similarly implying that ICC is exactly the same for two classes (De Beer, 2004). The shape of ICC for one class of examinees is thus below that of the other group in his opinion, as illustrated in Fig. 1.

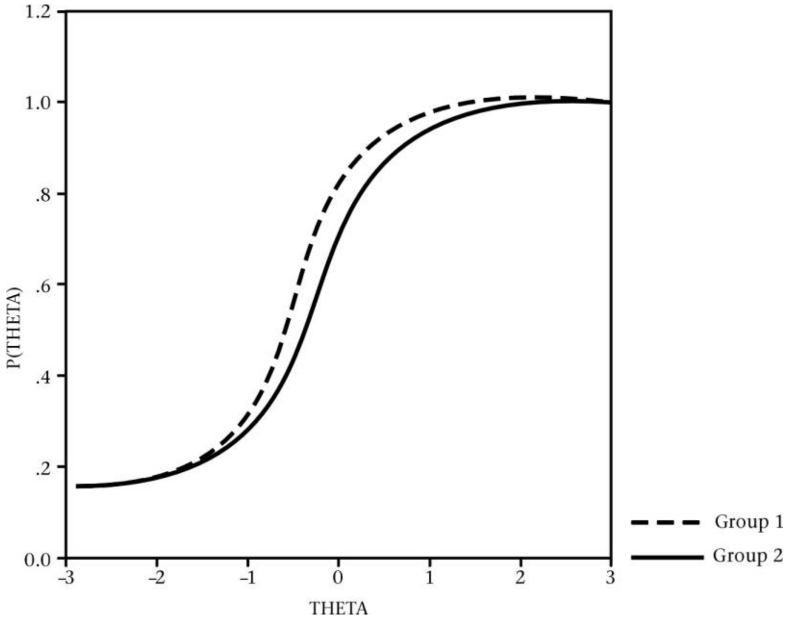

Figure 1. Item showing uniform DIF adopted from De Beer (2004).

Non-uniform DIF affects examinees inconsistently. When two groups are different on their slopes, the item is known to have non-uniform DIF. In other words, ICCs have different shapes for a different group of examinees in the case of non-uniform DIF. De Beer (2004) states that ICC shapes cross at a given point denoting that one group has a lesser possibility to answer the items accurately while such possibility for the other group is still higher. Fig. 2 illustrates the ICC shape for an item displaying the non-uniform DIF.

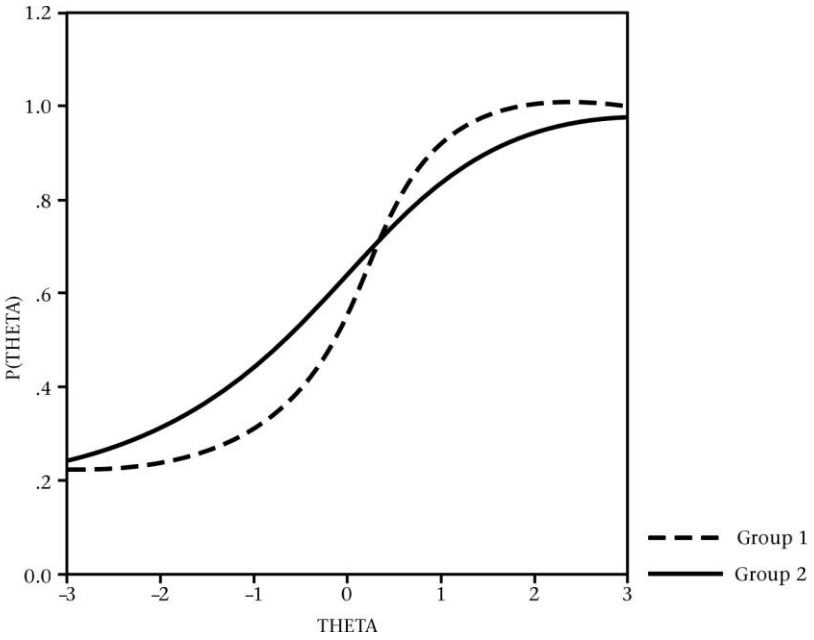

Figure 2. Item showing non-uniform DIF adopted from De Beer (2004).

De Beer (2004) claims the best and the most appropriate case is that there might be minimal differences between the shapes of ICC of comparing groups as depicted earlier.

2.4. The Empirical Perspectives on DIF

Chen and Henning (1985) investigated DIF for test-takers with various language backgrounds including Chinese and Spanish. They used Transformed Item Difficulty (TID) presented first by Angoff (1993). One hundred eleven test takers including 77 Chinese and 34 Spanish test takers took part in the research. Yet, the participants were not that much sizeable for the difficulty parameter to be reliably estimated. Lawrence et al.(1988) and Lawrence and Curley (1989) examined DIF as regards students' gender in the Scholastic Aptitude Test (SAT) using the standardization approach. The results indicated that females performed less well on items comparing to males. All these studies however comprised some drawbacks. First, most of them dealt with identifying DIF (uniform and non-uniform) with regard to item discrimination. Second, most studies that have been done on comparing the students’ total scores by means of standardization procedures showed that items were not usually purified before DIF detection. This may endanger the results of the studies. Ownby and Waldrop-Valverde (2013) applied IRT to determine whether the way the participants reacted to the items had any impact on older readers in a cloze test. They could detect 24 items flagging DIF. They concluded that DIF was a substantial source of variance that may endanger test scores’ interpretations and uses. Koo (2014) carried out meta-analytic DIF analyses on a reading test and the Florida Comprehensive Achievement Test (FCAT) by taking language, gender, and ethnicity into account. He realized that items having to do with vocabulary and phraseology favored non-English language learners regardless of their gender and ethnicity. Aryadoust and Zhang (2015) used the Rasch Model to test reading comprehension in a Chinese context. They found that while class one performed better on vocabulary, grammar, and general English proficiency, class two outperformed on skimming and scanning parts. The findings of the previous studies confirmed that gender had a slight impact on the individuals’ performance (Chen & Jiao, 2014; Hong & Min, 2007). Federer,et al (2016) investigated the correlation between the way male and female participants adopt while answering the open-ended questions. They found that women performed better under novel circumstances. In another study focusing on evolution, Smith (2016) made instrumentation dealing with the Evolution Theory. He could succeed to make a distinction between high school and university students utilizing items flagging DIF.

2.5. The Present Study

Owen (1992) states that the main motive behind conducting any research does not need to rely solely upon identifying and eliminating biased items but recognizing the elements contributing to bias is still more important. The present paper aimed at identifying the items that were susceptible to DIF as well as determining the group (subject fields) which was advantaged in those items. Meantime, most DIF investigations have been based upon the comparisons between gender (e.g., Ahmadi & Darabi Bazvand, 2016; Ahmadi & Jalili, 2014; Carlton, 1992; Federer et al., 2016; Karami, 2011; Lawrence et al., 1988), ethnicity (Koo, 2014; Schmitt, 1990), and language (Chen & Henning, 1985; Ryan

& Bachman, 1992) to-date. There are few studies that examined DIF for students with different subject fields. Thus DIF detection for students with different subject fields (Humanities vs. sciences) would be worth investigating. The main objective of this article was to detect questions displaying DIF on the MSRT (MCHE) proficiency test for test-takers with different fields of study (Humanities vs. Sciences) through IRT analysis. To the end, one research question motivated this study:

RQ**.** Do different subject fields (Humanities vs. Sciences) have any significant impact on Iranian EFL learners' performance? In other words, do test items function differently for test-takers with different fields of study (Humanities vs. Sciences)?

3. Methodology

3.1. Design and Context of the Study

Considering the fact the researchers couldn't maneuverer and control the independent variables, the design of this study was ex post facto as already confirmed by Hatch and Farhady (1982). Such design is normally used when there is no interference on part of the researchers on the participants' traits. This study comprised the test-takers' subject fields as an independent variable and their MSRT (MCHE) test scores as the dependent variable.

3.2. Participants

This study included two hundred pre-intermediates to intermediate EFL learners (100 female and 100 male students). The age range of these students was between 25 to 32. They were Ph.D. applicants as well as students of the doctorate in two different fields of study (100 Humanities and 100 sciences) in Tehran. All the participants spoke Persian as their L1. Table 1 demonstrates the demographic data of the participants:

Table 1.

Demographic Background of the Participants

No. of Students	200
Gender	(100 males- 100 Females)
Native Language	Persian
Major	EFL
Universities	Different Universities
Academic Year	2014-2015

3.3. Instruments

Parallel with the purposes of the article, the researchers applied one instrument as follows:

3.3.1. Ministry of Science, Research, and Technology (MSRT)

This test came into existence first in 1992 by the Ministry of Culture and Higher Education (MCHE) to check the proficiency level of Ph.D. candidates and/or students in various majors. In the year 2009, by changing the name of the above ministry to the Ministry of Science, Research and Technology, the name of the test was changed to MSRT. Since there have been a lot of students who participated in the exam before the year 2009, the authorities decided to mention both acronyms to avoid any inconveniences for the students and universities. At present, it is recognized as MSRT (MCHE) test. The criterion for acceptance is different in each university. In other words, each university has determined its own criterion. This test surprisingly lacks speaking skills. The MSRT (MCHE) proficiency exam consists of three parts: listening comprehension, grammar and structure, and reading comprehension. Each part is briefly summarized hereunder.

I. The Listening Comprehension

This section consists of 30 items as below:

a. In the first part, the test-takers need to listen to a brief part and choose the correct answers.

b. In the second part, the test-takers listen to some brief conversations between two people and then they have to choose the best answers; and

c. The third part presents some brief dialogues with different themes and asks test-takers to choose the correct answers.

II. The Structure and Written Expression

This section contains 30 items to assess the examinees' abilities to recognize the accurate English structures in two levels;

a. The first level needs test-takers to read an incomplete sentence and choose the word or a phrase that best completes it; and

b. The second level includes items with several words underlined in a sentence. The test-takers need to choose the one that consists of an unacceptable English.

III. The Reading Comprehension

This section comprises 40 items and is designed to measure test-takers' abilities to understand word meanings and reading-related materials. It contains some reading texts with different lengths and topics. The test-takers need to go through the questions and answer them thoroughly.

3.4. Data Collection Procedures

The researchers requested the MSRT training department to provide the raw scores for each section as well as the total scores of some participants. Upon the researchers' written request and by having met some administrative formalities, MSRT staff represented the test results of a large number of examinees that had already taken the test on the same day. The scores for each part were estimated based on the correct responses and no negative marks were considered for wrong answers. Once all the required data were collected, they were entered into SPSS v. 24 program. Then, the analyses were conducted through 3PL IRT Model. During the administration of the MSRT (MCHE) test, the usual precautions were met:

- Strict administration procedures were followed to minimize the effects of external factors like cheating, etc.
- Participants were not allowed to have anything other than the test papers on their desks.
- Participants were not allowed to take notes or make marks on their test papers.
- Participants were not permitted to complete any part of the exam before or after the given time.

3.5. Data Analysis Procedures

The scored items of two hundred Iranian EFL test-takers were entered into the IRT 3PL Model implying the likelihood that a test taker with an ability of theta (θ) responds to an item accurately, as regards item difficulty (b parameter), item discrimination (a parameter), and pseudo-guessing (c parameter) (Hambleton, Swaminathan, & Rogers, 1991). These features are mathematically shown hereunder:

$$P\ (\theta, a, b, c) = c + (1 - c)\frac{\exp(a(\theta - b))}{1 + \exp(a\theta - b))}$$

Where θ is the estimated ability, *a* is item discrimination, *b* is item difficulty, and *c* is the pseudo-guessing parameter. Since the *c* parameter is often poorly estimated, a prior distribution (*M* = 0.2 and *SD* = 1, according to Thissen (1991) has been applied. Thissen, Steinberg, and Wainer (1988) proposed that prior speculation is imposed on the *c* parameters when DIF is investigated using the 3PL IRT model. The IRT LR is a model-based approach that compares a compact model where all parameters are constrained to be equal across groups, hence no DIF, with an augmented model, can be detected. The fit of each model to the data is evaluated using the likelihood ratio goodness-of-fit statistic, G^2, and statistical difference in G^2 between the two models were also tested based on the chi-square statistics. Then, item discrimination (i.e., *a* parameter), item difficulty (i.e., *b* parameter), and G^2 were estimated employing likelihood ratio chi-square statistics. If *a* parameter is stable, it shows uniform DIF or no DIF. If the result is significant (i.e., variant *b* parameter), it indicates uniform DIF, and if not significant (i.e., invariant *b* parameter), no DIF. On the other hand, if *a* parameter of the studied item is not invariant, it demonstrates the presence of non-uniform DIF regardless of the *b* parameters.

4. Results

4.1. Section of Listening Comprehension (LC)

This part including 30 items was analyzed with respect to 3PL IRT Model to detect items flagging DIF. As Thissen, Steinberg, and Wainer (1988) confirmed, the effects of the *c* parameter were controlled in advance. Table 2 demonstrates the results. As it is shown in Table 2, six items (4, 6, 7, 13, 17, and 29) were identified to show DIF at the 0.05 significance level. Two items (i.e., items 7 and 17) displayed no DIF, and four items (i.e., items 4, 6, 13, and 29) exhibited non-uniform DIF.

Table 2

Listening Comprehension (LC)

Item	b	a	C	G2	X2	P
1	33.5%	.09	25%	1.82	1.81	.17
2	27.5%	.14	25%	.62	.62	.42
3	42.5%	.25	25%	.51	.51	.47

Item	b	a	C	G2	X2	P
4	32.5%	.30	25%	5.16	5.12	**.02**
5	3.5%	.03	25%	.14	.14	.70
6	48%	.38	25%	6.52	6.49	**.01**
7	43%	.25	25%	4.01	3.99	**.04**
8	44.5%	.25	25%	.02	.02	.88
9	44%	.14	25%	.35	.32	.56
10	32%	.12	25%	.36	.36	.54
11	56.5%	.18	25%	.99	.99	.31
12	48%	.21	25%	.08	.08	.77
13	23.5%	.18	25%	4.79	4.70	**.03**
14	36.5%	.07	25%	1.75	1.74	.18
15	31.5%	.23	25%	.58	.57	.44
16	38.5%	.21	25%	.02	.02	.88
17	43%	.25	25%	5.24	5.22	**.02**
18	33.5%	.20	25%	2.72	2.71	.09
19	17.5%	.16	25%	.31	.31	.57
20	36%	.14	25%	.34	.34	.55
21	27%	.23	25%	.10	.10	.75
22	28%	.16	25%	.89	.89	.34
23	17.5%	.16	25%	.03	.03	.85
24	21.5%	.20	25%	2.41	2.40	.12
25	12.5%	.07	25%	.41	.41	.52
26	17%	.05	25%	2.28	2.26	.13
27	11.5%	.03	25%	2.44	2.40	.12
28	11.5%	.09	25%	.04	.04	.82
29	12.5%	0	25%	8.07	7.72	**.00**
30	10.5%	.07	25%	.48	.47	.48

4.2. Structure and Written Expression (SWE)

This part included 30 items. To detect/identify DIF, each item was scrutinized concerning 3PL IRT model. The probable effects of the *c* parameter were controlled in advance, as recommended by Thissen et al. (1988). Table 3 summarizes the results. As Table 3 illustrates, five items (40, 43, 44, 45, and 57) were identified to show DIF at the 0.05 significance level.

Table 3

Structure and Written Expression (SWE)

Item	b	a	C	G2	X2	P
31	31.5%	.21	25%	1.13	1.13	.28
32	36%	.03	25%	.34	.34	.55
33	51.5%	.09	25%	3.39	3.38	.06
34	35%	.09	25%	1.40	1.40	.23
35	49.5%	.07	25%	.18	.18	.67
36	59%	.05	25%	.08	.08	.77
37	44.5%	.20	25%	.99	.99	.31
38	48%	.14	25%	.08	.08	.77
39	59.5%	.12	25%	.02	.02	.88
40	62%	.27	25%	6.92	6.87	**.00**
41	59.5%	.23	25%	.51	.51	.47
42	58%	.18	25%	.32	.32	.56
43	49	.29	25%	13.68	13.52	**.00**
44	65	.38	25%	4.33	4.30	**.03**
45	39	.36	25%	6.85	6.81	**.00**
46	40.5	.32	25%	.02	.02	.88
47	56.5	.43	25%	.50	.50	.47
48	49.5	.56	25%	.18	.18	.67
49	27	.29	25%	.91	.91	.33
50	43	.40	25%	.08	.08	.77
51	35.5	.41	25%	.02	.02	.88
52	25	.36	25%	.10	.10	.74
53	37.5	.47	25%	.19	.19	.66
54	36	.47	25%	.34	.34	.55
55	23	.29	25%	1.01	1.01	.31
56	36	.45	25%	2.17	2.17	.14
57	31	.48	25%	9.47	9.35	**.00**
58	30.5	.38	25%	.59	.59	.44
59	21.5	.34	25%	.26	.26	.60
60	23.5	.23	25%	3.39	3.36	.06

4.3. Reading Comprehension (RC)

This section included 40 items. To detect/identify DIF, each item was investigated with respect to the 3PL IRT model while the probable effects of the *c* parameter were controlled in advance as per Thissen, Steinberg, and Wainer's (1988) recommendations. Table 4 illustrates the results. As it is shown in Table 4, four items (61, 74, 80, and 97) were identified to show DIF at the 0.05 significance level.

Table 4

Reading Comprehension (RC)

Item	**b**	**a**	**C**	**G2**	**X2**	**P**
61	20.5	.25	25%	5.25	5.18	**.02**
62	26.5	.16	25%	.64	.64	.42
63	32	.20	25%	2.30	2.29	.13
64	15	.12	25%	1.42	1.41	.23
65	38	.16	25%	1.36	1.35	.24
66	30.5	.10	25%	.21	.21	.64
67	23	.09	25%	.00	.00	1.00
68	37.5	.12	25%	.02	.02	.88
69	30	.12	25%	.58	.85	.35
70	39	.12	25%	.08	.08	.77
71	49.5	.20	25%	.02	.02	.88
72	44.5	.21	25%	1.64	1.64	.20
73	47	.25	25%	2.89	2.89	.08
74	42	.16	25%	5.28	5.25	**.02**
75	50	.20	25%	.00	.00	1.00
76	45.5	.34	25%	2.44	2.44	.11
77	52.5	.10	25%	.02	.02	.88
78	51	.23	25%	.72	.72	.39
79	63.5	.27	25%	.54	.53	.46
80	47.5	.50	25%	4.52	4.51	**.03**
81	73	.32	25%	.91	.91	.33
82	59.5	.25	25%	.51	.51	.47
83	63.5	.38	25%	1.75	1.74	.18
84	64.5	.27	25%	.19	.19	.65
85	53	.40	25%	1.28	1.28	.25
86	48.5	.40	25%	.50	.50	.47
87	34.5	.40	25%	.02	.02	.88
88	61	.41	25%	1.34	1.34	.24
89	63	.25	25%	.08	.08	.77
90	35.5	.41	25%	3.70	3.69	.05
91	43	.29	25%	.73	.73	.39
92	33	.30	25%	.36	.36	.54
93	57	.27	25%	.73	.73	.39
94	55	.10	25%	1.29	1.29	.25
95	42.5	.23	25%	.18	.18	.66
96	34.5	.25	25%	.19	.19	.65
97	29	.34	25%	4.79	4.76	**.02**
98	27.5	.21	25%	.02	.02	.87
99	26.5	.12	25%	.23	.23	.63
100	29.5	.10	25%	1.49	1.49	.22

4.4. Comparing Two Groups Based on Descriptive Statistics

To find out which group (Humanities vs. Sciences) performed better at the exam in each part and the whole test, the independent samples t-test for means comparison of two groups has been estimated. Table 5 depicts the descriptive statistics for the Comparison of two groups of test-takers in three skills. As Table 5 reveals, the mean score of Science test-takers in the LC section (*M=9.36*) is higher than the Humanities' (*M=8.53*). Regarding SWE, as illustrated in Table 5, the mean score of Science test-takers (*M=13.89*) is higher than the Humanities' (*M=11.69*). With reference to RC, as it is shown in Table 5, the mean score of Science test-takers (*M=18.44*) is higher than that of Humanities (*M=16.35*). As far as the Total test is concerned, as Table 5 shows, by considering the mean score of Science test takers and the standard deviation (*M=41.72*, *SD=10.10*) and comparing them with those of Humanities (*M=36.57*, *SD= 11.48*), it turned out that Science test-takers outperformed the Humanities.

Table 5

Descriptive Statistics for the Comparison of Two Groups (Humanities vs. Sciences) in Three Skills

Group		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Total score	Humanities	100	36.57	11.48	1.14
	Sciences	100	41.72	10.10	1.01
LC	Humanities	100	8.53	4.12	0.41
	Sciences	100	9.36	3.94	0.39
SWE	Humanities	100	11.69	4.82	0.48
	Sciences	100	13.89	4.11	0.41
RC	Humanities	100	16.35	5.29	0.52
	Sciences	100	18.44	5.29	0.52

As Table 6 reveals, the mean score difference of Science and Humanities test-takers in the LC is not significant at the 0.05 level. Regarding SWE, as illustrated in Table 6, the mean score difference of Science and Humanities test-takers is significant at the 0.05 level. With reference to RC, as it is shown in Table 6, the mean score difference of Science and Humanities test-takers is not significant at the 0.05 level. As far as the Total test is concerned, as Table 6 shows, by considering the mean score and the standard deviation differences of Science and Humanities test-takers, it can be inferred that the exam was statistically easier for Science test-takers at 0.05 level.

Table 6

Independent Sample t-Test for Comparing Two Groups (Humanities vs. Sciences) in Each Part of the Exam and the Whole Test

		Levene's Test For equality of variances		**t-test for equality of means**		
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**
Total score		0.238	0.62	-3.36	198	0.00
Equal variances assumed				-3.36	194.85	0.00
	Equal variances not- assumed					
LC	Equal variances assumed	0.005	0.94	-1.45	198	0.14
	Equal variances not-assumed			-1.45	197.58	0.14
SWE	Equal variances assumed	2.399	0.12	-3.47	198	0.00
	Equal variances not-assumed			-3.47	193.18	0.00
RC	Equal variances assumed	0.006	0.93	-2.79	198	0.00
	Equal variances not-assumed			-2.79	198.00	0.00

The mean scores for test takers' performance on the whole test have been illustrated in Figure 3.

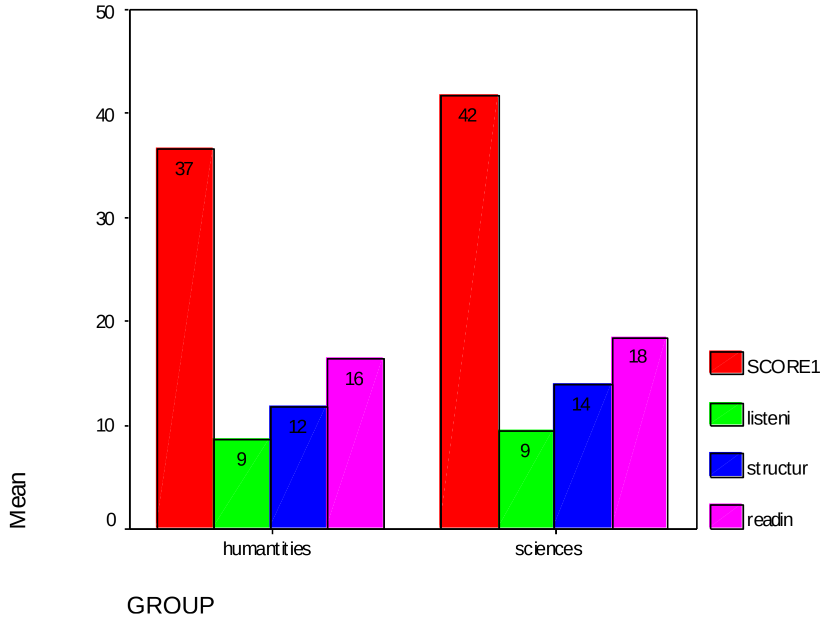

Figure 3. Mean scores of two groups (Humanities vs. Sciences).

Meanwhile, the raw score descriptive statistics and reliability estimates are also given in Table 7 for data sample (*n* = 200) results on the MSRT (MCHE) total test as well as its three sections. As displayed in Table 7, the MSRT (MCHE) Test has been proved to be a quite reliable test. The reliability for MSRT (MCHE) test as well as LC, SWE and RC parts were .85, .68, .72, and .73 respectively.

Table 7

Reliability Estimate Analyses

Skill		**Mean**	**Std. D.**	**N**	**R**
LC	Humanities	8.53	4.12	100	
	Sciences	9.36	3.94	100	.68
SWE	Humanities	11.69	4.82	100	
	Sciences	13.89	4.11	100	.72
RC	Humanities	16.35	5.29	100	
	Sciences	18.44	5.29	100	.73
Total	Humanities	36.57	11.48	100	
	Sciences	41.72	10.10	100	.85

5. Discussion

Identifying and eliminating DIF items are important for test fairness and validity. It is crucial to guarantee that latent traits of all test-takers were determined accurately by items and test scores. Although MSRT (MCHE) test has been undergone rigorous changes and revisions since its development, both test-takers and test-developers still doubt whether the test is fair for all groups of individuals. To address such ambiguities, the present study applied the IRT 3PL model to MSRT (MCHE) proficiency exam to discern items flagging DIF. The criterion variables were LC, SWE, RC, and examinees' academic major. Findings showed that items in different sections might be related to some features of individuals and may thus create bias in evaluating their proficiency in LC, SWE, and RC sections. However such discrepancies were not that much great. This denoted that the difficulty level of items was not the same for two groups of examinees in different fields of study.

As already confirmed by Zumbo (2007), these inconsistencies among examinees' performance may relate to some existing covariates. In this study, almost ten percent of the original 100 questions were eventually flagged as items displaying differential item functioning. They need to be discarded from the test's next administration. These findings contradict the general international findings proposed by McBride (1997). He believes one-third of the original items need to be removed in any test. Meantime, the findings of this research are in line with earlier studies where grammar, vocabulary, and reading comprehension were found to cause variations among examinees' performance and likely caused DIF (Grabe, 2009; Koda, 2005).

The findings of the study are inconsistent with those of Salehi and Tayebi (2012). They carried out the gender DIF study in reading comprehension of the Korean National Entrance Exam for Colleges and Universities. The findings showed that none of the 35 items of the reading examinees regarding their gender. They concluded that the test was fair to all male and female test-takers. However, the outcomes are in line with Ahmadi and Jalili's (2014) study. They applied two DIF detection methods of LR and IRT across an Iranian reading comprehension test. They found that 17% of the items displayed DIF, suggesting that item types such as reference and vocabulary were better predictors of gender DIF (mostly favoring females) than test content. The findings are also consistent with Ahmadi and Darabi Bazvand's (2016) research. They investigated gender differential item functioning across the Ph.D. Entrance Exam of TEFL (PEET) in an Iranian context. Using both logistic regression (LR) and one-parameter item response theory (1-p IRT) models, they concluded that PEET suffered from DIF as they identified 12% (12 items) of the whole test flagging for DIF with equal numbers of items showing uniform (six items) and non-uniform DIF (six items).

Tittle (1982) and Clauser (1990) suggest such items might cause the subgroup to be less motivated on the exam. At the same time, there are other unknown/different sources that may cause DIF. Since DIF is usually examined when comparing various groups of students is concerned, a big DIF value shows the existence of another construct that may lead to the differences/distinctions among the test takers. In sum, it is strongly recommended that the test-developers apply DIF analysis as an important part of their programs and be assured of the quality and content of their tests prior to administration (Zandi, et al., 2014) to enhance the assessment procedures. Integrating statistical analysis with the researchers' expertise might help understand whether DIF flagged items are fair or not.

6. Conclusion

With respect to the outcomes of this study, it can be inferred that once data was partitioned by variable under study, different permutations of variables emerged. In this study, thirteen out of a hundred items have been detected as items flagging DIF. As a general finding, those test-takers whose academic major was science outperformed the humanities students, especially in SWE and RC sections. SWE and RC play significant parts in any language proficiency test and are thus significant to devote further time and energy in a learning context to teach these parts more thoroughly. Learners should be aided to have a better appreciation of the significance and importance of these factors and do their best to improve in these areas. This study has some implications for MSRT (MCHE) test developers and those who take the test. The former is recommended either to carry out more studies to detect the items that may flag DIF or take care of the researchers' findings in this regard, and the latter can be assured the test scores are not favored against any particular type of examinees. Nevertheless, since gender is also a contributing factor, it is suggested to carry out a post hoc study to examine the effect of gender variable and identify the items that cause DIF due to that variable. Another point that is suggested for future studies is to consider how other variables including participants' background knowledge, test wise-ness, L1, culture, etc. would reveal more information about the items displaying DIF. The IRT model allows the researchers to gain access to a perceptible description of bias that is easy to understand and interpret. The findings of this study may help test developers to recognize sources of bias. It is essential to remind that test developers' ultimate interests may place in the type of decisions that are made according to test scores since test takers' circumstances depend upon such decisions in the future either partially or impartially. Recent methods in psychometric analysis are suggested to be developed and applied in further studies as new innovations might allow the researchers to carry out empirical investigations and may increase the precision of measurement.

References

Ahmadi, A., & Darabi Bazvand, A. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research, 4*(1), 63-82.

Ahmadi, A., & Jalili, T. (2014). A confirmatory study of Differential Item Functioning on EFL reading comprehension. *Applied Research on English Language, 3*(6), 55-68.

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*. Lawrence Erlbaum.

Aryadoust, V., & Zhang, L. (2015). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing, 33*(4), 529-553. https://doi:10.1177/0265532215594640.

Baghaei, P. (2008). RASCH measurement. Transactions of the Rasch measurement SIG. *American Educational Research Association, 22*(1), 1145-1146.

Baker, F. B. (1985). *The basics of item response theory*. Heinemann.

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.

Chen, Y.-F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. *Educational Assessment, 19*, 77-96.

Chen, Z., & Henning, G. . (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*, 155–163.

Clauser, B. E., & Mazor, K.M. . (1990). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31–47.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology, 30*(4), 52-58.

Ertuby, C., & Russel, R.J.H. . (1996). *Dealing with comparability problem of cross-cultural data.* Paper presented at the 26th International Congress of Psychology, Montreal.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531-1540.

Federer, M. R., Nehm, R. H., & Pearl, D. K. (2016). Examining gender differences in written assessment tasks in biology: A case study of evolutionary explanations. *CBE Life Sciences Education, 15*(1), ar2-ar2. https://doi:10.1187/cbe.14-01-0018

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*.-Cambridge University Press.

Hambleton, R. K. (1996). Guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment*.

Hambleton, R. K., Swaminathan, H., & Rogers, H.J. . (1991). *Fundamentals of item response theory*.-Sage Publications.

Hatch, E. M., & Farhady, H. . (1982). *Research design and statistics for applied linguistics*. –Rahnama Publications.

Hong, S., & Min, S.-Y. (2007). Mixed Rasch modeling of the self-rating depression scale incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement - EDUC PSYCHOL MEAS*, *67*, 280-299. doi:10.1177/0013164406292072.

Hope, D., Adamson, K., McManus, I.C. et al. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC Medical Education, 18*(64), 1143-0.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5*(2), 167-178.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.

Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing*, *31*(1), 89-109.

Lawrence, I. M., & Curley, W.E. (1989). Differential item functioning for males and females on SAT verbal reading subscore items: follow-up study. *Educational Testing Service Research Report*, 89–22.

Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). *Differential item functioning for males and females on SAT verbal reading subscore items*. College Entrance Examination Board.

Li, Z., & Zumbo, B.D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicologica, 30*(2), 343-370.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*.-Lawrence Erlbaum Assoc.

Mantel, N., & Haenszel, M.W. (1959). Statistical aspects of three analysis of data from retrospective studies of disease. *J Nat Cancer Inst., 22*, 719-748.

McBride, J. R. (1997). Technical perspective. *American Psychological Association*, 29-44.

Mousavi, A., & Krishnan, V. (2016). Measurement invariance of early development instrument (EDI) domain scores across gender and ESL status. *Alberta Journal of Educational Research*, *62*(3), 288-305.

Oliveri, M.E., Lawless, R., Robin, F., & Bridgeman, B. (2018). An exploratory Analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education, 31*(1), 1-16.

Osterlind, S. J. (1983). *Test item bias*. Sage.

Owen, K. (1992). *Test-item bias: Methods, findings and recommendations*. Human Sciences Research Council.

Owen, K. (1998). *The Role of psychological tests in education in South Africa [microform] : Issues, Controversies and Benefits / K. Owen*. Distributed by ERIC Clearinghouse.

Ownby, R. L., & Waldrop-Valverde, D. (2013). Differential item functioning related to age in the reading subtest of the test of functional health literacy in adults. *Journal of Aging Research, 2013*. https://doi:10.1155/2013/654589.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. https://doi:10.1177/014662169001400208.

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.

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12-29. https://doi:10.1177/026553229200900103.

Salehi, M., & Tayebi, A. (2012). Differential item functioning (DIF) in terms of gender in the reading comprehension subtest of a high-stakes test. *Iranian Journal of Applied Language Studies, 4*(1), 135-168.

Schmitt, A., & Dorans, N. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*, 67–81.

Smith, M. U., Snyder, S. W., & Devereaux, R. S. . (2016). The GAENE—generalized acceptance of evolution evaluation: Development of a new measure of evolution acceptance. *Journal of Research in Science Teaching, 53*, 1289–1315.

Thissen, D. (1991). *MULTILOG (Version 6.30) [Computer Software]*. Scientific Software.

Thissen, D., Steinberg, L., & Wainer, H. . (1988). *Use of item response theory in the study of group differences in trace lines.* Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1994). *Detection of differential item functioning using the parameters of item response models*. Lawrence Erlbaum.

Tittle, C. K. (1982). *Use of judgmental methods in item bias studies*. Johns Hopkins University Press.

Van de Vijver, F. (1998). Multicultural assessment: How suitable are western tests? *European Journal of Psychological Assessment, 14*(1), 61.

Zandi, H., Kaivanpanah, SH., & Alavi, S.M. (2014). The effect of test specifications review on improving the quality of a test. *Iranian Journal of Language Teaching Research 2*(1), 1-14.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233. https://doi:10.1080/15434300701375832