Applied & Educational Psychology (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/appep.2022.030103
ISSN 2523-5842 Vol. 3 Num. 1

Assessing Differential Item Functioning (DIF) For Pearson Test of English (PTE) A study of Test Takers with Different Fields of Study

Hamed Ghaemi[1,a,*]

[1]*Bahar Institute of Higher Education, Mashhad, Iran*
ghaemiacademy@gmail.com
**corresponding author*

Keywords: Differential Item Functioning (DIF); Item Response Theory (IRT); Likelihood Ratio Approach (LR); Fields of Study, Pearson Test of English.

Abstract: Differential Item Functioning (DIF), which is a statistical feature of an item and provides a sign of unpredicted performance of items on a test, occurs once dissimilar groups of test takers with the same level of ability show different performance on a single test. The aim of this paper was to examine DIF on the Pearson Test of English (PTE) test items. To that end, 250 intermediate EFL learners with the age range of 26 - 36 in two different fields of study (125, Engineering, and 125 Sciences) were randomly chosen for the analysis. The Item Response Theory (IRT) Likelihood Ratio (LR) approach was utilized to find items showing DIF. The scored items of 250 PTE test takers were analyzed using the IRT three-parameter model utilizing item difficulty (b parameter), item discrimination (a parameter), and pseudo-guessing (c parameter). The results of the independent samples t-test for comparison of means in two groups depicted that Science participants performed better than the Engineering ones particularly in Speaking & Writing and Reading sections. It is evident that the PTE test was statistically easier for the Science students at 0.05 level. Linguistic analyses of Differential Item Functioning items also confirmed the findings of the quantitative part, indicating a far better performance on the part of Science students.

1. Introduction

The growth of the psychometric tests and testing procedures have been affected by virtue of social and political fluctuations within the few past decades (Owen, 1998). When psychometric tests are used to perform individual or group comparisons, item bias ought to be considered to lessen the unfitting interpretations. Test bias varies from test fairness in that it is usually measured quantitatively while test fairness is carried out subjectively and intuitively and it is not feasible to be described in absolute terms, indicating that no one can categorize tests as either fair or not fair. It can be taken as read that it is not the test characteristics being significant on its own but the scores' interpretations and the results that are of overriding significance as the students' educational future is usually determined by these decisions. The term *biased* pertains to the applied instruments, testing procedures and the methods of scores interpretation. The scores' differences between two groups don't merely

define the term bias (Osterlind, 1983). The term bias has been superseded by differential item functioning (DIF) showing that individuals who are parallel considering their level of ability have different performance on a test and gain various scores accordingly. Test bias or DIF is concerned with systematic errors and discloses the characteristics associating with item psychometric characteristics depicting that the items cannot measure impartially considering different individuals/groups.

In actual fact, DIF arises when "individuals from various classes have the similar ability level but display different likelihood in responding to an item accurately" (Osterlind, 1983, p. 32). Basically, non-DIF represents the situation in which the test takers with the analogous level of ability irrespective of their in-group differences have the same probability to answer an item correctly. DIF deals with the extent to which the test items differentiate between participants having the same ability level from various groups consisting of gender, ethnicity, education, etc. (Zumbo, 2007). Parameters contributing to item/test bias are "culture, education, language, socioeconomic status, and so on" (Van de Vijver, 1998, p. 35). Test bias or DIF should be evaluated and calculated during test construction process (Osterlind, 1983). Tests ought to be constructed in a way that when inconsistency in examinees' test results is observed, such discrepancy is attributed to differences in the construct that the test is going to assess. By detecting and eliminating items demonstrating DIF as well as the analysis of items, test developers find problematic items lacking psychometric properties. This paper investigated item analysis of PTE, an internationally – recognized proficiency test, by means of item response theory (IRT) based on DIF study.

2. Literature Review

2.1. Methods of DIF identification

Finding items demonstrating DIF permits the test developers to match the examinees with the pertinent knowledge. DIF is concerned with the students' scores on the tests, their hidden ability's measurement and examination of individuals being analogous with reference to their level of capability and come from various background though perform identical on an item. *Mantel- Haenszel* x^2 *Test* is used for detecting DIF (Mantel and Haenszel, 1959), suiting well even for small number of participants and empowers the test makers to utilize simple arithmetic measures based upon logistic regression methods proposed by Zumbo (2007). Modest arithmetic procedures offer a more in-depth explanation of DIF and permits the researchers to make distinction between uniform and non-uniform DIF. The other procedures to detect DIF employ IRT models as stated by Lord, (1980), Raju (1990), and Thissen, Steinberg, & Wainer (1994). These methods deal with examinees' ability and characteristics of items more accurately and are more concerned with larger sample sizes. Among these models, IRT is used more by the researchers to spot items flagging DIF, as these models "render the most useful data for identifying differences on particular items" (Ertuby, 1996, p. 51).

2.2. Models of Item Response Theory (IRT)

Most of the measurement procedures, in particular in the field of education and psychology, deal with the latent variables (Hambleton, 1996). The chance of answering correctly hinge on both item characteristics and examinees' level of ability. Such a relationship is mathematically stated as item characteristic curve (ICC). Any ICC ought to envisage the examinees' scores based on their underlying abilities, which is also recognized as item response function. The examinees' level of abilities is shown along the X-axis and represented by theta (θ) while the probability of responding to items correctly is demonstrated on Y-axis and is shown by p (θ). As Baker (1985) proposed, the ICC shape rest on the item difficulty (b-parameter), item discrimination (a-parameter), and guessing

power known as pseudo-chance (c-parameter). In fact, depending on horizontal location, ICCs might vary, spotting the individuals' ability level against items' difficulty. The likelihood of selecting 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, ranging from -2.5 to +2.5 in theory. Meaning it differs from the very easy items to very tough ones.

Item discrimination (a-parameter) displays the slope of the ICC and the accuracy of the measurement of a given item. The curve slope and item discrimination are positively correlated in a 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 larger discrimination power may well differentiate the individuals. The guessing power (c-parameter) displays the probability a test taker with the bottommost level of ability answering the item accurately. The c-parameter ranges from 0 to 1. IRT models alter concerning the properties of items they involve. The one parameter or Rasch model has to do with the item difficulty and ability level of examinees. The two parameter model deals with item discrimination and Item difficulty (probability of getting the correct response based on examinees' ability level). Third parameter or pseudo-chance parameter is realized when items have multiple-choice format and 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 Non-uniform vs. Uniform DIF

DIF usually has two distinct categories with regard to logistic regression model: uniform and non-uniform. Uniform DIF affects the participants at all levels equally suggesting that ICC is precisely identical for two classes. De Beer (2004) believes that the likelihood to pick the correct answer is less than that of another class in uniform DIF. The shape of ICC for one class of testees is therefore below that of the other group in his opinion, as illustrated in Fig. 1.

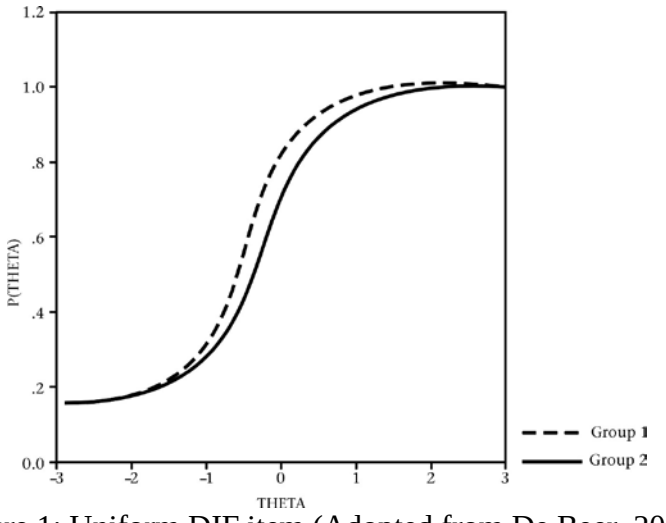

Figure 1: Uniform DIF item (Adopted from De Beer, 2004)

When two groups are different on their slopes, the item shows non-uniform DIF. In other words, ICCs have various shapes for different groups of examinees in non-uniform DIF. Non-uniform DIF influences examinees inconsistently. De Beer (2004, p. 42) states that "the ICC shapes cross at a given point implying that one group has a lesser possibility to answer the test items accurately while such possibility for the other group was still higher". Fig. 2 shows the ICC shape for an item demonstrating the non-uniform DIF.

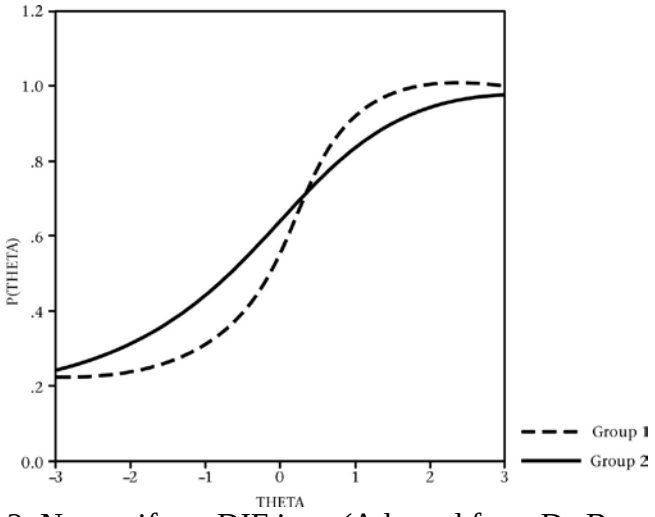

Figure 2: Non-uniform DIF item (Adopted from De Beer, 2004)

A DIF analysis for test takers with various language backgrounds encompassing Chinese and Spanish was examined by Chen and Henning (1985). They employed Transformed Item Difficulty (TID) presented first by Angoff (1993). TID provides the item difficulty indices between two groups of test takers and identified outliers. One hundred eleven test takers including seventy-seven Chinese and thirty-four Spanish test takers took part in the research. Nevertheless, the participants were not that much ample for the difficulty parameter to be consistently measured. Lawrence, Curley, & McHale (1988) and Lawrence & Curley (1989) studied DIF regarding students' gender in the Scholastic Aptitude Test (SAT) by dint of the standardization method. The outcomes depicted females performed not as well on items as male test takers. All these studies, though, have some downsides. First, most of them dealt with finding DIF (uniform and non-uniform) considering item discrimination. Furthermore, most studies conducted on comparing the students' total scores through standardization processes have shown that items are not typically examined before DIF detection. This may jeopardize the results of the studies. Ownby and Waldrop-Valverde (2013) applied IRT to determine whether the way the participants respond to the items has any influence on older readers in a cloze test. They spotted twenty four items flagging DIF, concluding that DIF was a substantial cause of variance that may imperil test scores' interpretations and uses. Koo (2014) conducted meta-analytic DIF analyses on a reading test and the Florida Comprehensive Achievement Test (FCAT) by taking language, gender, and ethnicity into account. He figured out that items of vocabulary and phraseology favored non-English language learners irrespective of their gender and ethnicity. Aryadoust and Zhang (2015) utilized a Rasch model to a test of reading comprehension in a Chinese context. They found that while class one performed better on vocabulary, grammar, and general English proficiency, the other class surpassed in skimming and scanning parts. The results of most prior studies showed the gender had a trivial impact on the performance of the readers (Hong & Min, 2007; Chen & Jiao, 2014). Federer, Nehm, & Pearl (2016) explored the correlation between the way male and female participants while answering the open-ended questions. They found that women performed better under novel circumstances. In another study focusing on evolution, Smith (2016) made an instrumentation dealing with the Evolution Theory. He could succeed to make a distinction between high school and university students using items flagging DIF.

3. The Current Study

The present paper aimed at finding and identifying the items that were susceptible to DIF as well as determining the fields of study which were advantaged in those items. Most DIF investigations are based upon the comparisons between gender (e.g., Lawrence, Curley, & McHale, 1988; Carlton, 1992; Federer et al., 2016), ethnicity (Schmitt, 1990; Koo, 2014), or language (Chen & Henning, 1985; Ryan & Bachman, 1992) to-date. There are insufficient studies which scrutinized DIF for students with different subject fields focusing on PTE as an international proficiency test. Thus, DIF detection for students with different subject fields (Engineering vs. Sciences) willing to participate in PTE, worth investigating. The main objective of this paper was to detect questions displaying DIF on PTE proficiency test for test takers with different fields of study (Engineering vs. Sciences) by means of IRT analysis. To the end, two research questions motivated this study:

RQ1: Do test items (PTE test) function differently for test takers with different fields of study (Engineering vs. Sciences)?

RQ2: Are there linguistic features of these items that account for the DIF results?

4. Methodology

4.1 Participants

This study included 250 intermediate EFL learners with the age range of 26 -36. They were Ph.D. applicants as well as Master's degree holders in two different fields of study (125 Engineering) and 125 Sciences) in Iran. All the participants spoke Persian / Farsi as their L1.

4.2 Instruments

In line with the purposes of the study, the researchers applied one instrument as follows:

4.2.1 Pearson Test of English (PTE)

Pearson Language Tests is devoted to measuring and validating the English language of non-native English speakers. The tests comprise the Pearson Test of English (PTE) Academic, PTE General and PTE Young Learners. These are administered in association with Edexcel, the world's largest examining body. In 2009, Pearson Language Tests introduced the Pearson Test of English Academic which is recognized by Graduate Management Admission Council (GMAC). The test score has been associated to the levels well-defined in the Common European Framework of Reference for Languages (CEFR). PTE Academic is distributed through the Pearson Virtual User Environment (VUE) centers which are also in charge of holding the GMAT (Graduate Management Admission Test). Upon publicizing, it was accepted by nearly 6,000 organizations. As a case in point, the test is accepted by the Australia Border Agency and the Australian Department of Immigration and Citizenship for visa applications. The test is mostly read by a computer rather than a human corrector to decrease waiting times of the results for students.

Table 1. Detailed pattern of PTE:

Part one: Speaking and Writing (70 – 90 minutes)				
Enabling Skills	**Main Skill**	**Scoring Method**	**Number of Questions**	Question Type

Oral fluency, pronunciation, Content	Reading & speaking	Partial credit	6 -7	Read aloud
Oral fluency, pronunciation, Content	Listening & speaking	Partial credit	10 - 12	Repeat Sentence
Oral fluency, pronunciation, Content	Speaking	Partial credit	6 - 7	Describe image
Listening and speaking Oral fluency, pronunciation Content	Writing	Partial credit	3 - 4	Re-tell lecture
Vocabulary	Listening & speaking	Correct/incorrect	10 -12	Answer short question
Grammar, vocabulary Content, form	Reading & Writing	Partial credit	2 - 3	Summarize written text
Grammar, vocabulary, spelling, written discourse Content; development, structure and coherence; form, general linguistic range	Writing	Partial credit	1 -2	Write essay

Part two: Reading (31 – 42 minutes)

Enabling Skills	**Main Skill**	**Scoring Method**	**Number of Questions**	Question Type
-	Reading	Correct/incorrect	2 -3	Multiple-choice, choose single answer
-	Reading	Partial credit (for each correct response. Points deducted for incorrect options chosen)	2 -3	Multiple-choice, choose multiple answers
-	Reading	Partial credit (for each correctly ordered, adjacent pair)	2 - 3	Re-order paragraphs
-	Reading	Partial credit (for each correctly completed blank)	4 -5	Reading: Fill in the blanks
-	Reading	Partial credit (for each correctly completed blank)	5 -6	Reading and writing: Fill in the blanks

Part three: Listening (45 – 57 minutes)

Enabling Skills	**Main Skill**	**Scoring Method**	**Number of Questions**	Question Type
Grammar, vocabulary, spelling Content, form	Listening & Writing	Partial credit	2 -3	Summarize spoken text
-	Listening	Partial credit (for each correct response. Points deducted for incorrect options chosen)	2 -3	Multiple–choice, choose multiple answers
-	Listening & Writing	Partial credit (each correct word spelled correctly)	2 -3	Fill in the blanks
-	Listening & Writing	Correct/ incorrect	2 -3	Highlight correct summary
-	Listening	Correct/ incorrect	2 -3	Multiple-choice,

				choose single answer
-	Listening	Correct/ incorrect	2 -3	Select missing word
-	Listening & Reading	Partial credit (for each word. Points deducted for incorrect options chosen)	2 -3	Highlight incorrect words
-	Listening & Writing	Partial credit (for each word spelled correctly)	3 - 4	Write from dictation

4.3 Data Collection Procedures

The researchers requested the PTE candidates to provide them with report card of their score in each section as well as the total scores. In addition to this, the scores of each item were collected and used for the purpose of data analysis. The scores for each part had been estimated based on the correct responses and no negative marks had been considered for wrong answers. During the administration of the PTE test, the usual precautions were met:

1. Strict administration procedures were followed to minimize the effects of external factors like cheating, etc.
2. For any form of ID to be acceptable it will need to be a valid document (not expired) or its issue date no more than 10 years old.
3. The same ID details shared when booking the test must be presented by the test taker on the day of the test.
4. The name on the ID should exactly match the name used when booking the test.
5. If you fail to produce the required ID you will not be allowed into the test room and will lose your test fee.
6. Copies will not be accepted. The original document must be provided. No other ID will be accepted at the test center.

4.4 Design

In view of the fact the researchers couldn't manipulate 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 utilized when there is no interference on the part of the researchers on the participants' traits. This study comprised the test-takers' subject fields as an independent variable and their PTE test scores as the dependent variable.

4.5 Data Analysis Procedures

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

$$P\,(x = 1/\theta) = c + \frac{1 - c}{1 + e^{-Da(\theta - b)}}$$

Where, x is an item response, θ is the estimated ability, a is item discrimination, b is item difficulty, c is pseudo-guessing parameter, D is a scaling factor (= 1.7) that is devised to estimate the IRT models to a cumulative normal curve, and e is a transcendental number whose value is 2.718.

However, because the *c* parameter is often poorly assessed, a prior distribution (M = 0.2 and SD = 1, according to Thissen (1991) has been applied. Thissen, Steinberg, & Wainer (1988) proposed that a prior speculation is applied to the *c* parameters when DIF is studied using the 3PL IRT model. The IRT LR is a model-based approach and compares a model in which all parameters are controlled to be equal across groups, hence no DIF, with an amplified model, permitting parameters to be free across groups. Using the likelihood ratio goodness-of-fit statistic, G^2, the fit of each model to the data is estimated. 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 measured by means of probability ratio of chi-square statistics. If *a* parameter is constant, it confirms unchanging, uniform DIF or no DIF. If the result is significant (i.e., variant *b* parameter), it designates uniform DIF. On the other hand, if *a* parameter of the studied items is variant, it proves the presence of non-uniform DIF in spite of the *b* parameters.

5. Results

5.1 The Outcomes of Research Question

The results of DIF investigations on IRT 3P LR model are shown in Tables 2, 3, and 4. These Tables depict the following data:

1. (b) standing for Item Difficulty
2. (a) standing for Item Discrimination
3. (c) revealing Guessing
4. (G2) revealing Likelihood Ratio Goodness-of-fit
5. (X2) representing Chi-square
6. (P) representing the Probability or Test of Significance

5.1.1. Speaking & Writing

This part included 38 - 57 items. To have clear understanding and detailed and reliable calculations, in this study 57 questions are considered for the *Speaking and Writing* part, which is the utmost number of items in PTE *Speaking and Writing* part. This is actually applied for other parts of the test too). To detect/identify DIF, each item was analyzed with respect to 3PL IRT model. To do this, as Thissen, Steinberg, & Wainer (1988) confirmed, the impacts of *c* parameter were controlled in advance. As it is shown in Table 2, twelve items (4, 6, 7, 13, 17, 29, 34, 38, 46, 47, 52 and 53) 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, 29, 50, 55 and 57) exhibited non-uniform DIF.

Table 2. Speaking and Writing

Item	b	a	C	G2	X2	P
1	24.5%	.07	25%	1.452	1.568	.188
2	27.5%	.16	25%	.748	.877	.418
3	43.5%	.23	25%	.412	.452	.494
4	34.5%	.31	25%	5.143	5.458	**.014**
5	4.5%	.04	25%	.219	.188	.701
6	38%	.33	25%	6.746	6.870	**.011**
7	53%	.24	25%	4.123	3.888	**.036**
8	54.5%	.26	25%	.040	.040	.877
9	54%	.13	25%	.252	.775	.529
10	36%	.19	25%	.268	.398	.554
11	54.5%	.14	25%	.878	.977	.358
12	47%	.27	25%	.090	.090	.797
13	27.5%	.17	25%	4.494	4.790	**.020**
14	34.5%	.09	25%	1.761	1.44	.136
15	30.5%	.24	25%	.850	.519	.477
16	37.5%	.23	25%	.041	.031	.894
17	41%	.27	25%	5.547	5.42	**.012**
18	37.5%	.21	25%	2.785	2.91	.089
19	14.5%	.11	25%	.232	.372	.567
20	37%	.13	25%	.457	.357	.576
21	28%	.24	25%	.151	.121	.760
22	25%	.19	25%	.874	.873	.355
23	18.5%	.17	25%	.055	.025	.892
24	23.5%	.21	25%	2.494	2.900	.181
25	13.5%	.09	25%	.742	.741	.531
26	18%	.04	25%	2.546	2.588	.172
27	13.5%	.05	25%	2.232	2.777	.191
28	12.5%	.03	25%	.069	.099	.885
29	14.5%	.19	25%	8.06	7.82	**.005**
30	11.5%	.08	25%	.478	.479	.479
31	17.5%	.09	25%	.419	.491	.469
32	17.2%	.06	25%	1.368	.596	.391
33	14.9%	.17	25%	.894	.674	.731
34	12.8%	.16	25%	.364	1.297	**.004**
35	23.2%	.26	25%	.477	.779	.611
36	12.9%	.03	25%	.335	.364	.574
37	19.9%	.14	25%	2.585	2.747	.331
38	13.3%	.08	25%	.894	1.775	**.002**
39	17.3%	.06	25%	.661	.771	.390
40	13.6%	.13	25%	.97	.987	.284
41	23.9%	.07	25%	.413	.651	.574
42	20.8%	.04	25%	1.247	1.749	.837
43	29.7%	.23	25%	.985	3.511	.462
44	13.9%	.18	25%	.689	.368	.378
45	43.6%	.09	25%	.371	.746	.567
46	18.9%	.06	25%	1.657	.654	**.030**
47	19.7%	.22	25%	2.965	1.105	**.014**
48	41.8%	.23	25%	2.329	4.149	.268
49	36.9%	.15	25%	.357	3.364	.964
50	15.7%	.17	25%	.374	.301	.775
51	11.9%	.08	25%	.952	.357	.394
52	13.8%	.06	25%	.635	.741	**.020**

53	28.9%	.08	25%	.478	.459	**.002**
54	12.9%	.09	25%	.598	.201	.584
55	13.7%	.11	25%	.365	.988	.137
56	10.9%	.18	25%	.321	.740	.791
57	10.5%	.08	25%	.854	1.204	.594

5.1.2. Reading

This part included 20 items. To detect/identify DIF, each item was scrutinized with respect to 3PL IRT model. The plausible effects of *c* parameter were controlled in advance, as recommended by Thissen, Steinberg, & Wainer (1988). As Table 3 indicates, five items (3, 10, 13, 14 and 15) were found to depict DIF at the 0.05 significance level.

Table 3. Reading

Item	b	a	C	G2	X2	P
1	33.5%	.22	25%	1.122	1.665	.288
2	39%	.02	25%	.334	.365	.598
3	57.5%	.07	25%	3.553	3.433	**.033**
4	36%	.07	25%	1.659	1.437	.265
5	48.5%	.08	25%	.131	.187	.673
6	58%	.06	25%	.088	.098	.763
7	45.5%	.22	25%	.987	.966	.365
8	49%	.13	25%	.087	.090	.788
9	56.5%	.14	25%	.073	.033	.875
10	66%	.26	25%	6.986	6.536	**.009**
11	58.5%	.27	25%	.576	.553	.454
12	59%	.16	25%	.355	.356	.543
13	45%	.25	25%	13.77	13.72	**.000**
14	66%	.37	25%	4.344	4.448	**.028**
15	36%	.37	25%	6.67	6.88	**.009**
16	41.5%	.38	25%	.029	.022	.822
17	54.5%	.44	25%	.588	.566	.411
18	47.5%	.52	25%	.132	.126	.621
19	28%	.22	25%	.954	.966	.331
20	43%	.42	25%	.033	.087	.771

5.1.3 Listening

This section includes 25 items. To detect/identify DIF, each item was investigated with respect to 3PL IRT model while the probable effects of *c* parameter were controlled in advance as per Thissen, Steinberg, & Wainer's (1988) recommendations. As it is shown in Table 4, four items (1, 14, and 20) were recognized to show DIF at the 0.05 significance level.

Table 4. Listening

Item	b	a	C	G2	X2	P
1	21.5%	.26	25%	5.35	5.22	**.013**
2	23.5%	.15	25%	.623	.641	.333
3	31%	.22	25%	2.316	2.958	.140
4	16%	.13	25%	1.210	1.552	.225
5	34%	.15	25%	1.230	1.36	.254
6	31.5%	.11	25%	.211	.214	.665
7	21%	.02	25%	.000	.000	1.10
8	39.5%	.12	25%	.022	.022	.894
9	31%	.13	25%	.577	.851	.395
10	38%	.17	25%	.088	.074	.792
11	45.5%	.21	25%	.021	.010	.898
12	46.5%	.20	25%	1.952	1.850	.210
13	48%	.22	25%	2.628	2.888	.090
14	43%	.13	25%	5.580	5.241	**.012**
15	51%	.21	25%	.000	.000	1.10
16	42.5%	.30	25%	2.665	2.421	.198
17	53.5%	.13	25%	.021	.021	.877
18	57%	.22	25%	.787	.721	.376
19	66.5%	.26	25%	.521	.535	.493
20	46.5%	.51	25%	4.329	4.881	**.024**
21	72%	.30	25%	.914	.995	.399
22	56.5%	.23	25%	.545	.574	.491
23	66.5%	.36	25%	1.881	1.957	.196
24	68.5%	.23	25%	.199	.199	.688
25	56%	.46	25%	1.856	1.545	.297

5.1.4 Comparing two groups based on Descriptive Statistics

To discover which group (Engineering vs. Sciences) performed better at the exam in each part and the whole test, the independent samples t-test for comparison of means in two groups has been carried out. As Tables 5 and 6 illustrates, the mean score of Science test takers in Listening section (10.36) is higher than the Engineering test takers (9.33). However, the difference is not significant at 0.05 level. Regarding Speaking and Writing (S & W), as shown in Tables 5 and 6, the mean score of Science test takers (14.89) is higher than the Engineering' (10.69). Such difference is significant at 0.05 level. Concerning Reading, as it is illustrated in Tables 5 and 6, the mean score of Science test takers (19.94) is higher than that of Engineering (15.55). However, the distinction is not significant at 0.05 level. As for the Total test, as Tables 5 and 6 demonstrate, by considering the mean score of Science test takers (45.52) and the standard deviation (SD=11.11) and comparing them with those of Engineering (35.55); (SD= 13.38), it turned out that Science test takers outperformed the Engineering. It can be inferred that the exam was statistically easier for Science test takers at 0.05 level.

Table 5. Descriptive statistics for the Comparison of Two Groups (Engineering vs. Sciences) in Three Parts of PTE

Group	N	Mean	Std. Deviation	Std. Error Mean
Total score Engineering	125	35.5500	13.38143	1.14814
Sciences	125	45.5200	11.11558	1.01056
Listening Engineering	125	9.3300	4.12789	0.41279
Sciences	125	10.3600	3.94282	0.39428

S & W	Engineering	125	10.6900	4.82145	0.48214
	Sciences	125	14.8900	4.11181	0.41118
Reading	Engineering	125	15.5500	5.29031	0.52903
	Sciences	125	19.9400	5.29593	0.52959

Table 6. Independent sample t-test for comparing two groups (Engineering vs. Sciences) in each part of the exam as well as 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	Equal variances assumed	0.338	0.716	-4.457	198	0.001
	Equal variances not-assumed			-4.457	196.869	0.001
Listening	Equal variances assumed	0.006	0.855	-1.744	198	0.148
	Equal variances not-assumed			-1.764	187.565	0.148
S & W	Equal variances assumed	2.469	0.213	-3.562	198	0.001
	Equal variances not-assumed			-3.562	197.145	0.001
Reading	Equal variances assumed	0.007	0.926	-2.782	198	0.007
	Equal variances not-assumed			-2.782	198.000	0.007

In the meantime, the descriptive statistics and reliability estimates are also given in Table 7 for data sample (n = 250) results on the PTE total test as well as its three sections. As presented in Table 7, the PTE Test has been proved to be a quite reliable test. The reliability for the whole PTE test as well as Listening, Speaking & Writing and Reading parts were .95, .88, .82 and .93 respectively.

Table 7. Reliability Estimate Analyses

Skill	**Fields of Study**	**N**	**R**
Listening	Engineering	125	.88
	Sciences	125	
S & W	Engineering	125	.82
	Sciences	125	
Reading	Engineering	125	.93
	Sciences	125	
Total	Engineering	125	.95
	Sciences	125	

6. Qualitative results

6.1. Linguistic analyses of differential item functioning items

In order to have a better understanding of the results from the DIF analyses, the researcher undertook an investigation of the linguistic features of these items. The goal of this linguistic analysis was to determine whether the DIF findings between pairs of fields of study could be explained by the information of variances across study fields in one of the following:

- The linguistic differences or similarities between the two fields of study, i.e. Engineering and Science.
- The approaches or methods that are frequently utilized to teach English in the two fields of study.

Consistent with previously – mentioned procedures for the DIF analyses, the results from the linguistic analyses are organized first by section of the assessment, then by pairwise fields of study analyses within each section.

6.1.1. Speaking and Writing Part

Eight of the items that favored the Science test-takers, 17, 29, 34, 38, 46, 47, 52 and 53, were items that tested vocabulary usage and Fluency in speaking. For these items, all of the questions had their roots in Latin, which tended to favor the Science students, because most of their university textbooks are written in English. Also, science students are claimed to have much better speaking skill since their classes at university are delivered in English. In fact, the medium of instruction in General English Classes of Science students is English, favoring their speaking proficiency. Four items exhibited C-level DIF, 4, 6, 7 and 13, all of which favored the Engineering test-takers. All of these items tested grammatical usage and may have favored the Engineering students due to curriculum and/or instructional differences experienced by the Engineering students. In fact, their university General English book mainly focused on various grammatical features.

6.1.2. Reading Comprehension

Five items exhibited C-level DIF, 3, 10, 13, 14 and 15. Three of the items, 13, 14 and 15 favored the Science test-takers, while two items, 3 and 10 favored the Engineering test-takers. There are several possible explanations for why the Science students performed better on the items that exhibited DIF in their favor. On average, the Science test-takers performed better than the Engineering students on the Reading Comprehension section of PTE. Items 3 and 10 both required students to make a higher-level inference from their reading, a task in which the Science test-takers may have been stronger than the Engineering students. Item 13 was concerned about the meaning of a phrase, in which Science students performed far better. Item 14 required test-takers to make an inference based on an indirect phrasing; again, this may have been a task in which the Science test-takers may have been stronger than the Engineering students. Item 15 was cognitively more challenging, favored the Engineering test- takers who were older than the Science test-takers and were more likely to own the high-level cognitive skills required to recognize the implied main idea of a passage.

6.1.3. Listening Part

Item 1 and 14 favored the Science test-takers, while Item 20 favored the Engineering test-takers. Item 1 and 14 tested lower-level listening skills, precisely skills in remembering information. Item 20 tested higher level listening skills, particularly, being able to categorize the implied main idea of the listening passage. That is, Item 1 and 14 favored the Science test-takers because they are mostly sturdier in lower-level cognitive skills (remember, understand, and apply), while Item 20 favored the Engineering test-takers, who were older than the Science test-takers and may have been more likely to possess the high-level cognitive skills (concept acquisition, systematic decision making, evaluative thinking, brainstorming) required to detect the implicit main idea.

7. Discussion

Finding and removing DIF items are significant for test fairness and validity. It's vital to guarantee that latent traits of all test-takers are determined precisely by items and test scores. Although PTE test has experienced severe vicissitudes 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 obscurities, the present study applied IRT 3PL model to PTE proficiency exam to distinguish items flagging DIF. The criterion variables were *Listening, Speaking & Writing, Reading,* and examinees' academic field of study. Findings depict that items in different parts might be associated to some features of individuals and may therefore create bias in assessing their proficiency. Nevertheless, such inconsistencies were not that much great, denoting 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 discrepancies among examinees' performance may be linked to some prevailing covariates. In this study, almost twenty percent of the original questions ultimately flagged as items showing differential item functioning. They need to be discarded from the test's next administration. These findings oppose with the general international results proposed by McBride (1997). He believes one third of original items needs to be deleted in any test. The findings of this research are in line with earlier studies where speaking, vocabulary, listening and reading were found to cause disparities among examinees' performance and caused DIF (Grabe, 2009; Koda, 2005). Tittle (1982) and Clauser (1990) suggest such items might cause the target group to be less inspired on the exam. Simultaneously, there are other unknown sources that may cause DIF. In light of the fact that DIF is usually scrutinized when comparing innumerable groups of students is concerned, a big DIF value illustrates the presence of extra construct that may lead to the alterations/distinctions among the test takers. All in all, it is highly recommended that the test-developers utilize DIF analysis as a significant aspect of their programs to augment the assessment procedures. Mixing statistical analysis with the researchers' knowledge and skills might help the test developers realize whether DIF tagged items are fair or not.

8. Conclusion

With regard to the findings of this study, it can be taken as read that when data was divided according to variable under study, different variations of variables arose. In this study, twenty out of 91 items have been detected as items flagging DIF. As a general finding, those test takers whose academic major was Science outperformed the Engineering students especially in S&W and Reading sections. S&W and Reading play pivotal role in any language proficiency test and are therefore substantial to dedicate further time and energy in learning context to teach these parts more systematically. Learners should be assisted to have a better appreciation of the implication and importance of these factors and do their best to ameliorate in these skills. This study has some implications for PTE test developers and those who take the test. The former are highly recommended to conduct more studies to identify the items that may flag DIF and take care of the researchers' findings in this regard, and the latter can be guaranteed that the test scores are not favored against any specific type of examinees. Nonetheless, given that the gender is also a contributing factor, it is recommended to perform a post hoc study to inspect the influence of gender variable and find the items that cause DIF owing to that variable. Another point worth suggesting for future studies is to contemplate how other variables consisting of participants' background knowledge, test wise-ness, L1, culture, etc. would disclose more information about the items showing DIF. The IRT model permits the researchers to access to a noticeable explanation of bias that is convenient to realize and construe. The outcomes of this study help test-developers to distinguish sources of bias. It is vital to recap that test developers' decisive interests may place in the kind of decisions that are made based

on test's scores as test takers' conditions depend upon such verdicts in future either partially or impartially. Recent methods in psychometric analysis are proposed to be established and applied in further studies as new novelties might permit the researchers to do experimental investigations and it may upsurge the accuracy of measurement.

References

[1] Angoff, W. H. (1993). Perspectives on Differential Item Functioning methodology. NJ: Lawrence Erlbaum.
[2] 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. doi:10.1177/0265532215594640
[3] Baker, F. B. (1985). The basics of item response theory. NH: Heinemann.
[4] 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.
[5] 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.
[6] Chen, Z., & Henning, G. . (1985). Linguistic and cultural bias in language proficiency tests. Language Testing, 2, 155–163.
[7] Clauser, B. E., & Mazor, K.M. . (1990). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice, 17, 31–47.
[8] 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.
[9] 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.
[10] 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. doi:10.1187/cbe.14-01-0018
[11] Grabe, W. (2009). Reading in a second language: Moving from theory to practice. Cambridge: Cambridge University Press.
[12] Hambleton, R. K. (1996). Guidelines for Adapting Educational and Psychological Tests European Journal of Psychological Assessment.
[13] Hambleton, R. K., Swaminathan, H., & Rogers, H.J. . (1991). Fundamentals of item response theory. CA: Sage Publications.
[14] Hatch, E. M., & Farhady, H. . (1982). Research design and statistics for applied linguistics.Tehran Rahnama Publications.
[15] Hong, S., & Min, S.-Y. (2007). Mixed Rasch Modeling of the Self-Rating Depression ScaleIncorporating Latent Class and Rasch Rating Scale Models. Educational and Psychological Measurement - EDUC PSYCHOL MEAS, 67, 280-299. doi:10.1177/0013164406292072
[16] Koda, K. (2005). Insights into second language reading: A cross-linguistic approach. New York: Cambridge University Press.
[17] Koo, J., Becker, B. J., & Kim, Y. S. (2014). 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.
[18] Lawrence, I. M., & Curley, W.E. (1989). (1989). Differential Item Functioning for males and females on SAT-Verbal Reading subscore items: follow-up study. Educational Testing Service Research Report, 89–22.
[19] Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). Differential item functioning for males and females on SAT verbal reading subscore items. New York: College Entrance Examination Board.
[20] Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. NJ.: Lawrence Erlbaum Assoc.
[21] Mantel, N., & Haenszel, M.W. (1959). Statistical aspects of thee analysis of data from retrospective studies of disease. J Nat Cancer Inst., 22, 719-748.
[22] McBride, J. R. (1997). Technical Perspective. American Psychological Association, 29-44.
[23] Osterlind, S. J. (1983). Test item bias. Beverly Hills: Sage.
[24] Owen, K. (1998). The Role of Psychological Tests in Education in South Africa [microform]: Issues, Controversies and Benefits / K. Owen. [Washington D.C.]: Distributed by ERIC Clearinghouse.
[25] 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, 654589-654589. doi:10.1155/2013/654589

[26] 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. doi:10.1177/014662169001400208
[27] Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. Language Testing, 9(1), 12-29. doi:10.1177/026553229200900103
[28] Schmitt, A., & Dorans, N. . (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27, 67–81.
[29] 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.
[30] Thissen, D. (1991). MULTILOG (Version 6.30) [Computer Software]. Chicago, IL: Scientific Software.
[31] Thissen, D., Steinberg, L., & Wainer, H. (1994). Detection of differential item functioning using the parameters of item response models. NJ: Lawrence Erlbaum.
[32] Thissen, D., Steinberg, L., & Wainer, H. . (1988). Use of item response theory in the study of group differences in trace lines. NJ: Erlbaum.
[33] Tittle, C. K. (1982). Use of judgmental methods in item bias studies. MD: Johns Hopkins University Press.
[34] Van de Vijver, F. (1998). Multicultural assessment: How suitable are Western tests? European Journal of Psychological Assessment, 14(1), 61.
[35] 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. doi:10.1080/15434300701375832.