Jurnal Penelitian dan Evaluasi Pendidikan
Volume 24, No 2, December 2020 (190-197)
Online: http://journal.uny.ac.id/index.php/jpep

PERFORMANCE DIFFERENCES BY GENDER IN ENGLISH READING TEST

Ari Arifin Danuwijaya[1], Adiyo Roebiyanto[2*]
[1]Department of English Education, Universitas Pendidikan Indonesia
Jl. Dr. Setiabudi No.229, Isola, Sukasari, Kota Bandung, Jawa Barat 40154, Indonesia
[2]Department of Psychology, Universitas Mercu Buana Jakarta
Jl. Raya Meruya Selatan No.1, Kembangan, Jakarta Barat 11650, Indonesia
*Corresponding Author. E-mail: adiyo.roebianto@mercubuana.ac.id

ABSTRACT
Test fairness becomes an aspect that needs to be considered when developing a test instrument. It is highly recommended that the instrument should not be biased for the test takers by ensuring that they do not behave differently among male and female test-takers. This study aims to examine the extent to which the items in an English proficiency test function differently across gender. Fifty reading items were examined and analyzed using a statistical method for detecting DIF. The items were individually tested for gender DIF using Rasch model analysis with the analysis tool of ConQuest. The results showed that six items were detected for DIF, three of which were basic comprehension items, and the other three were vocabulary questions. Some possible ways of dealing with DIF items were also discussed.
Keywords: *DIF, gender differences, test fairness, reading test, ConQuest*

How to cite: Danuwijaya, A., & Roebianto, A. (2020). Performance differences by gender in English reading test. *Jurnal Penelitian dan Evaluasi Pendidikan, 24*(2), 190-197. doi:https://doi.org/10.21831/pep.v24i2.34344

INTRODUCTION

It has been a commonplace for a single form of a test to have many different types of items to measure skill, knowledge or abilities. When the test items are administered to examinees, there is a potential to have items that function differently in some contexts that might favor one group of examinees. The presence of item DIF or potential bias with regards to different demographic characteristics, such as gender, social class, ethnicity, should be examined further to avoid bias and to promote test fairness (Huff, 2000; Kunnan, 2007; Le, 2006; Wu, Tam, & Jen, 2016). One way to ensure test fairness for the test takers is by understanding possible gender differences in the test items. The primary purpose of the study is to examine the test items that behave differently for different gender group. From the test taker responses, the study investigates the way items function differently for individuals or groups of test takers who have similar abilities (Kunnan, 1990).

According to Wu, Tam, and Jen (2016), the term Differential Item Functioning or DIF relates to an item that functions differently to different groups or contexts. DIF could be a factor that might affect test performance in favor of a particular group, such as gender. The groups might be gender, cultural background, geography, and ethnicity. An item which is exhibits DIF in gender group, for example, functions differently showing one group of male students performs considerably better than the other group on the item. As DIF might affect test performance in favor of one particular group (Takala & Kaftandjieva, 2000), it is essential to treat the items that are detected for DIF to make sure the test validity and fairness for the groups (Lin & Wu, 2003).

ISSN: 2685-7111 (print)
2338-6061 (online)

https://doi.org/10.21831/pep.v24i2.34344

A number of studies which investigated DIF have been well documented on some fields of areas, such as mathematics (Kan & Bulut, 2014; Ong, Williams, & Lamprianou, 2015) and second/foreign language testing (Kunnan, 1990; Pae, 2012; Zumbo, 2003). In one study, Kim and Jang (2009) investigated a number of reading items on the Ontario Secondary School Literacy Test (OSSLT) that function differentially for L1 students and ELL students. The findings showed that vocabulary knowledge items favored L1 students, while ELL students were favored by grammatical knowledge or integrating reading and writing skill. Kunnan (1990) examined the ESL placement examination (ESLPE) at the University of California, LA (UCLA) to investigate the identification of DIF among four native language groups and two gender groups. Using the one-parameter Rasch model for Item Response Theory (IRT) to a sample of 884 non-native speaking students at UCLA, the results showed some items displayed DIF in the native language groups (thirteen items) and in the gender group analysis (twenty three items). In gender analysis, 20 items favored male group and the items are found in the test sections of listening (seven items), reading (four items), grammar (three items), vocabulary (four items), writing error detection (two items), and grammar (three items). The source of the DIF in listening and reading was the passages that related to business, culture, and engineering disciplines, which favored the male group. While the potential source of DIF for vocabulary test was the test-taker major field.

Another study related to DIF in gender was conducted by Takala and Kaftandjieva (2000). The study analyzed 40 multiple-choice English vocabulary items to 182 males and 293 females in the intermediate level of the Finnish Foreign Language Certification Examination. The results showed some items advantaged females (five items) and males (six items), and the items were excluded from the item bank as biased estimates of person parameters were produced. In addition, a study by Pae (2012) investigated the potential causes of gender DIF on a high stake national test over a long period of time.

Considerable attention has been paid on examining DIF in various language proficiency tests, many is known about the skill items function differently for gender group. Thus, this study aims to investigate to the extent to which the items in PTESOL function differently for male and female test takers. In the end, this study provides is to evaluate the test for better validity.

RESEARCH METHOD

Data

The data used in this study were item-level responses which were taken from PTESOL test. The PTESOL is an English proficiency test that is designed to provide information on the examinees' abilities in English subject. The test is developed by the Language Center of Universitas Pendidikan Indonesia (UPI), and its development is guided by a test specification. The test was administered for senior high school students and plays a role as part of English proficiency evidence for university admission or employment. The test consists of three sections, namely listening comprehension, structure and written expression, and reading comprehension. The English reading comprehension section was considered for this study. The items are based on a multiple-choice format with five four options. The items are categorized into three four categories: reading for main idea, reading for basic comprehension, inferencing, and vocabulary knowledge, as presented in Table 1. For this study, the data were in the form of students' responses on reading comprehension subtests. The responses of 1,067 (consisting of 411 males and 656 females) year 3 senior high school students were used, and the data were taken from 2016 test administration from three public schools in different cities (Bandung, Cimahi, and Garut) in Indonesia.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i2.34344

Table 1. The PTESOL Reading Subskills and the Corresponding Items

Reading Subskill	Description of Reading Subskill	Item	No of Items
Reading for Main Idea (RMI)	Ability to comprehend information in a text by skimming the information	1, 22, 33, 42	4
Reading for Basic comprehension (RBC)	Ability to comprehend explicit or implicit information in a text by scanning and locating stated information	2, 4, 6, 7, 9, 10, 11, 12, 16, 19, 21, 23, 25, 26, 29, 31, 32, 35, 39, 41, 46, 48, 49, 50	24
Inferencing (INF)	Ability to draw inferences about explicitly stated information by carefully attending to an author's purpose, attitude, tone, etc	15, 17, 18, 28, 34, 40, 43,	7
Vocabulary Knowledge (VOC)	ability to comprehend meanings of words and phrases used in the context of the test	3, 5, 8, 13, 14, 20, 24, 27, 30, 36, 37, 38, 44, 45, 47,	15

Analysis

The statistical method for detecting DIF used in this study was item response theory (IRT). The 50 reading comprehension items were individually tested for gender DIF using Rasch model analysis with the analysis tool of ConQuest (Adams & Wu, 2010a). ConQuest was selected because of its powerful tool for examining DIF, particularly to model interactions between item and gender. It describes the probability of correct responses to generalized items using an item main effect, a gender main effect, and an interaction between item and gender (Adams & Wu, 2010b).

The first procedure for analysis was to conduct the analysis of fit statistics which provided information about how well the pattern of the observed responses match with the modeled expectation (Lee-Ellis, 2009). The fit statistics discusses both person and item fit. Person fit examines how different the patterns of the responses of the examinees, which can create spuriously high or spuriously low test scores, with the specified response model (Karabatsos, 2003; Reise, 1990). The upper limit of the IMS range for a person used the cut-off value proposed by Curtis and Boman (2007) is around 1.60. Meanwhile, the acceptable ranges of item fit statistics are from 0.7 to 1.4 (Curtis & Boman, 2007).

To identify DIF, items flagged for DIF were indicated by chi-square value and the absolute DIF value taken from the logit difference between two groups. Some scholars propose different cut-off value of logit difference to detect the existence of DIF. Le (2006) suggests that items are flagged for DIF if the chi-square DIF test is significant at 0.01 level and its absolute DIF value is greater than 0.25 logit. Meanwhile, Wu, Tam, and Jen (2016) suggests 0.5 logit as a cut-off value, while Bond and Fox (2015) propose a difference of 0.5 logit for high-stakes test. For the purpose of this analysis, the cut-off value of 0.5 was used to detect DIF.

FINDINGS AND DISCUSSION

As one of the advantages using IRT is to be able to detect and identify misfitting persons (Dodeen, 2003), the person fit statistics were first calculated. In the test, unusual responses from examinees are sometimes found and such responses create misfitting to the testing model. According to Dodeen (2003), misfitting is the source of inaccuracy in estimating an individual's ability and decreasing the test validity. Using ConQuest (Adams & Wu, 2010a) as a tool of analysis to detect examinees' responses and to identify those who misfit the model as well as using the cut-off value of 1.6, the findings showed that 104 misfitting persons were identified and removed from the analysis.

The second term yielded from ConQuest analysis showed the estimates of gender differences in ability estimates. One of the estimates value is in negative sign to indicate a group that performed poorly. According to Adams and Wu (2010b), the negative sign is used for the

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i2.34344

gender term in the item response model showing poor performance. The results show that the estimate value of male students is -0.05 and that of female students is 0.05 with the standard error of 0.03. It indicates that male students perform poorer than the female students. It is shown that the male students scored 0.10 lower than female students. The parameter estimate is more than its standard error and the chi-square p-value is 1.67, showing a lower value than 2 which indicates that the difference is not significant.

The third term based on ConQuest analysis gives information about the interaction between the item and gender facets (see Figure 1). Some items were found to be easy for male students than female students, and vice versa. Item 1, for example, had the estimate of 0.014, and indicating that 0.014 must be added to the difficulty of this item for male students, and -0.014 must be added for the females. This item is also relatively easier for female students than male students. Another example is item 5 with the estimate of -0.050 for males and 0.050 for females, indicating that male students found it easier compared to female students. Based on the table of parameter estimates (see Table 2), there are twenty four items (items5, 6, 9, 11, 14, 16, 19, 21-28, 30, 33, 34, 36, 40, 42, 46, 47, and 49) that are relatively easier for male students, and twenty six (items 1-4, 7, 8, 10, 12, 13, 15, 17, 18, 20, 29, 31, 32, 35, 37-39, 41, 43, 44, 45, 48, and 50) that are relatively easier for female students.

Table 2. Response Model Parameter Estimates

Item	Gender		Estimate	Error	Unweighted			Weighted			DIF
					MNSQ	CI	T	MNSQ	CI	T	Magnitude
1	1	male	0.014	0.105	0.91	(0.89, 1.11)	-1.5	0.97	(0.82, 1.18)	-0.3	0.028
2	1	male	0.094	0.083	0.94	(0.89, 1.11)	-1.1	0.95	(0.91, 1.09)	-1	0.188
3	1	male	0.051	0.071	0.94	(0.89, 1.11)	-1	0.95	(0.95, 1.05)	-2	0.102
4	1	male	0.077	0.071	1.04	(0.89, 1.11)	0.7	1.04	(0.94, 1.06)	1.3	0.154
5	1	male	-0.05	0.07	0.92	(0.89, 1.11)	-1.4	0.93	(0.95, 1.05)	-2.8	-0.1
6	1	male	-0.109	0.081	0.87	(0.89, 1.11)	-2.4	0.93	(0.89, 1.11)	-1.2	-0.218
7	1	male	0.333	0.086	0.82	(0.89, 1.11)	-3.3	0.91	(0.84, 1.16)	-1.2	0.666
8	1	male	0.002	0.089	1.44	(0.89, 1.11)	6.7	1.16	(0.86, 1.14)	2.2	0.004
9	1	male	-0.032	0.147	0.79	(0.89, 1.11)	-3.9	0.96	(0.69, 1.31)	-0.2	-0.064
10	1	male	0.128	0.101	1.1	(0.89, 1.11)	1.7	1.02	(0.80, 1.20)	0.2	0.256
11	1	male	-0.194	0.074	1.26	(0.89, 1.11)	4.2	1.21	(0.94, 1.06)	6.3	-0.388
12	1	male	0.051	0.07	1.08	(0.89, 1.11)	1.3	1.06	(0.95, 1.05)	2.6	0.102
13	1	male	0.299	0.083	1.11	(0.89, 1.11)	1.9	1.04	(0.86, 1.14)	0.6	0.598
14	1	male	-0.01	0.134	0.74	(0.89, 1.11)	-4.8	0.95	(0.73, 1.27)	-0.4	-0.02
15	1	male	0.176	0.074	0.87	(0.89, 1.11)	-2.3	0.9	(0.94, 1.06)	-3.7	0.352
16	1	male	-0.178	0.117	0.9	(0.89, 1.11)	-1.8	0.95	(0.75, 1.25)	-0.4	-0.356
17	1	male	0.179	0.112	1.04	(0.89, 1.11)	0.7	1.01	(0.83, 1.17)	0.2	0.358
18	1	male	0.116	0.071	1.02	(0.89, 1.11)	0.4	1.02	(0.95, 1.05)	0.6	0.232
19	1	male	-0.084	0.073	1.04	(0.89, 1.11)	0.7	1.04	(0.94, 1.06)	1.1	-0.168
20	1	male	0.31	0.078	0.94	(0.89, 1.11)	-1	0.96	(0.94, 1.06)	-1.4	0.62
21	1	male	-0.167	0.071	1.1	(0.89, 1.11)	1.6	1.09	(0.95, 1.05)	3.4	-0.334
22	1	male	-0.127	0.074	1.09	(0.89, 1.11)	1.6	1.04	(0.93, 1.07)	1.2	-0.254
23	1	male	-0.003	0.098	0.8	(0.89, 1.11)	-3.7	0.94	(0.84, 1.16)	-0.7	-0.006
24	1	male	-0.037	0.072	0.86	(0.89, 1.11)	-2.6	0.89	(0.94, 1.06)	-4.2	-0.074
25	1	male	-1.08	0.538	0.43	(0.89, 1.11)	-12.6	0.99	(0.00, 2.95)	0.3	-2.16
26	1	male	-0.034	0.073	0.89	(0.89, 1.11)	-1.9	0.91	(0.94, 1.06)	-3	-0.068
27	1	male	-0.072	0.07	1.05	(0.89, 1.11)	0.9	1.03	(0.95, 1.05)	1.4	-0.144
28	1	male	-0.101	0.075	0.97	(0.89, 1.11)	-0.4	0.99	(0.93, 1.07)	-0.2	-0.202
29	1	male	0.118	0.08	0.92	(0.89, 1.11)	-1.4	0.96	(0.92, 1.08)	-0.9	0.236
30	1	male	-0.167	0.071	1.06	(0.89, 1.11)	1	1.05	(0.95, 1.05)	2.1	-0.334
31	1	male	0.144	0.124	0.97	(0.89, 1.11)	-0.5	0.97	(0.79, 1.21)	-0.2	0.288
32	1	male	0.016	0.082	1.14	(0.89, 1.11)	2.3	1.07	(0.89, 1.11)	1.2	0.032
33	1	male	-0.093	0.079	1	(0.89, 1.11)	0	1.01	(0.91, 1.09)	0.3	-0.186
34	1	male	-0.101	0.081	1.07	(0.89, 1.11)	1.2	1.05	(0.89, 1.11)	0.8	-0.202
35	1	male	0.035	0.074	1.01	(0.89, 1.11)	0.3	1.01	(0.94, 1.06)	0.3	0.07
36	1	male	-0.08	0.071	1.09	(0.89, 1.11)	1.5	1.07	(0.95, 1.05)	2.8	-0.16
37	1	male	0.1	0.071	0.93	(0.89, 1.11)	-1.2	0.94	(0.94, 1.06)	-1.9	0.2
38	1	male	0.128	0.092	1	(0.89, 1.11)	0.1	1	(0.88, 1.12)	0	0.256
39	1	male	0.241	0.076	1.14	(0.89, 1.11)	2.4	1.1	(0.90, 1.10)	1.9	0.482
40	1	male	-0.102	0.08	0.97	(0.89, 1.11)	-0.6	0.99	(0.90, 1.10)	-0.3	-0.204

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i2.34344

Item	Gender		Estimate	Error	Unweighted			Weighted			DIF Magnitude
					MNSQ	CI	T	MNSQ	CI	T	
41	1	male	0.122	0.092	0.92	(0.89, 1.11)	-1.4	0.97	(0.88, 1.12)	-0.4	0.244
42	1	male	-0.093	0.088	1.06	(0.89, 1.11)	1.1	1.04	(0.88, 1.12)	0.6	-0.186
43	1	male	0.003	0.171	0.88	(0.89, 1.11)	-2.1	1	(0.63, 1.37)	0.1	0.006
44	1	male	0.227	0.072	1.03	(0.89, 1.11)	0.5	1.02	(0.95, 1.05)	0.9	0.454
45	1	male	0.32	0.186	1	(0.89, 1.11)	0	1.01	(0.48, 1.52)	0.1	0.64
46	1	male	-0.486	0.17	0.78	(0.89, 1.11)	-4	0.97	(0.52, 1.48)	0	-0.972
47	1	male	-0.045	0.094	0.87	(0.89, 1.11)	-2.3	0.94	(0.85, 1.15)	-0.7	-0.09
48	1	male	0.025	0.09	0.86	(0.89, 1.11)	-2.5	0.93	(0.87, 1.13)	-1.1	0.05
49	1	male	-0.085	0.143	0.97	(0.89, 1.11)	-0.6	1	(0.69, 1.31)	0.1	-0.17
50	1	male	0.222*	0.071	1.12	(0.89, 1.11)	2	1.09	(0.95, 1.05)	3.6	0.444
1	2	female	-0.014*	0.105	0.88	(0.86, 1.14)	-1.7	0.96	(0.76, 1.24)	-0.3	
2	2	female	-0.094*	0.083	0.89	(0.86, 1.14)	-1.5	0.94	(0.86, 1.14)	-0.9	
3	2	female	-0.051*	0.071	0.95	(0.86, 1.14)	-0.7	0.95	(0.94, 1.06)	-1.4	
4	2	female	-0.077*	0.071	0.96	(0.86, 1.14)	-0.6	0.97	(0.93, 1.07)	-0.9	
5	2	female	0.050*	0.07	0.97	(0.86, 1.14)	-0.4	0.97	(0.94, 1.06)	-0.8	
6	2	female	0.109*	0.081	0.8	(0.86, 1.14)	-2.9	0.89	(0.88, 1.12)	-1.8	
7	2	female	-0.333*	0.086	0.79	(0.86, 1.14)	-3.1	0.87	(0.88, 1.12)	-2.4	
8	2	female	-0.002*	0.089	1.39	(0.86, 1.14)	4.8	1.17	(0.84, 1.16)	2	
9	2	female	0.032*	0.147	0.95	(0.86, 1.14)	-0.6	0.97	(0.61, 1.39)	-0.1	
10	2	female	-0.128*	0.101	0.78	(0.86, 1.14)	-3.2	0.89	(0.81, 1.19)	-1.1	
11	2	female	0.194*	0.074	1.3	(0.86, 1.14)	3.8	1.22	(0.90, 1.10)	4.3	
12	2	female	-0.051*	0.07	1	(0.86, 1.14)	0.1	0.99	(0.94, 1.06)	-0.2	
13	2	female	-0.299*	0.083	1.23	(0.86, 1.14)	2.9	1.16	(0.89, 1.11)	2.9	
14	2	female	0.010*	0.134	0.77	(0.86, 1.14)	-3.3	0.95	(0.65, 1.35)	-0.2	
15	2	female	-0.176*	0.074	0.81	(0.86, 1.14)	-2.7	0.87	(0.90, 1.10)	-2.7	
16	2	female	0.178*	0.117	0.86	(0.86, 1.14)	-2	0.96	(0.74, 1.26)	-0.3	
17	2	female	-0.179*	0.112	1.01	(0.86, 1.14)	0.2	1.02	(0.72, 1.28)	0.2	
18	2	female	-0.116*	0.071	0.99	(0.86, 1.14)	-0.1	0.99	(0.93, 1.07)	-0.3	
19	2	female	0.084*	0.073	1.05	(0.86, 1.14)	0.7	1.05	(0.92, 1.08)	1.3	
20	2	female	-0.310*	0.078	0.86	(0.86, 1.14)	-2	0.93	(0.88, 1.12)	-1.2	
21	2	female	0.167*	0.071	1.09	(0.86, 1.14)	1.3	1.06	(0.93, 1.07)	1.5	
22	2	female	0.127*	0.074	1.15	(0.86, 1.14)	2	1.11	(0.92, 1.08)	2.4	
23	2	female	0.003*	0.098	0.82	(0.86, 1.14)	-2.5	0.96	(0.79, 1.21)	-0.4	
24	2	female	0.037*	0.072	0.83	(0.86, 1.14)	-2.4	0.87	(0.93, 1.07)	-3.6	
25	2	female	1.080*	0.538	0.67	(0.86, 1.14)	-5	0.99	(0.16, 1.84)	0.1	
26	2	female	0.034*	0.073	0.89	(0.86, 1.14)	-1.5	0.92	(0.92, 1.08)	-2.1	
27	2	female	0.072*	0.07	1.02	(0.86, 1.14)	0.3	1.02	(0.94, 1.06)	0.6	
28	2	female	0.101*	0.075	1.01	(0.86, 1.14)	0.1	1.01	(0.91, 1.09)	0.1	
29	2	female	-0.118*	0.08	0.87	(0.86, 1.14)	-1.9	0.94	(0.87, 1.13)	-0.8	
30	2	female	0.167*	0.071	1.24	(0.86, 1.14)	3	1.18	(0.93, 1.07)	4.8	
31	2	female	-0.144*	0.124	0.9	(0.86, 1.14)	-1.4	0.97	(0.67, 1.33)	-0.1	
32	2	female	-0.016*	0.082	1.24	(0.86, 1.14)	3.1	1.03	(0.87, 1.13)	0.5	
33	2	female	0.093*	0.079	0.98	(0.86, 1.14)	-0.2	1	(0.89, 1.11)	0.1	
34	2	female	0.101*	0.081	1.08	(0.86, 1.14)	1.1	1.05	(0.88, 1.12)	0.8	
35	2	female	-0.035*	0.074	0.92	(0.86, 1.14)	-1.1	0.95	(0.91, 1.09)	-1.2	
36	2	female	0.080*	0.071	1.04	(0.86, 1.14)	0.5	1.02	(0.94, 1.06)	0.6	
37	2	female	-0.100*	0.071	0.97	(0.86, 1.14)	-0.3	0.97	(0.94, 1.06)	-1.1	
38	2	female	-0.128*	0.092	0.86	(0.86, 1.14)	-1.9	0.94	(0.81, 1.19)	-0.6	
39	2	female	-0.241*	0.076	1.16	(0.86, 1.14)	2.1	1.12	(0.92, 1.08)	2.7	
40	2	female	0.102*	0.08	0.99	(0.86, 1.14)	-0.1	1	(0.88, 1.12)	0	
41	2	female	-0.122*	0.092	0.92	(0.86, 1.14)	-1.1	1.01	(0.81, 1.19)	0.1	
42	2	female	0.093*	0.088	0.99	(0.86, 1.14)	-0.2	1	(0.84, 1.16)	0	
43	2	female	-0.003*	0.171	0.52	(0.86, 1.14)	-8	0.91	(0.52, 1.48)	-0.3	
44	2	female	-0.227*	0.072	1.01	(0.86, 1.14)	0.2	1.01	(0.92, 1.08)	0.2	
45	2	female	-0.320*	0.186	1.07	(0.86, 1.14)	1	1.02	(0.57, 1.43)	0.2	
46	2	female	0.486*	0.17	0.92	(0.86, 1.14)	-1	0.94	(0.63, 1.37)	-0.3	
47	2	female	0.045*	0.094	0.79	(0.86, 1.14)	-3.1	0.93	(0.81, 1.19)	-0.7	
48	2	female	-0.025*	0.09	0.81	(0.86, 1.14)	-2.8	0.91	(0.83, 1.17)	-1.1	
49	2	female	0.085*	0.143	0.79	(0.86, 1.14)	-3	0.97	(0.63, 1.37)	-0.1	
50	2	female	-0.222*	0.071	1.09	(0.86, 1.14)	1.2	1.08	(0.93, 1.07)	2.3	

After misfitting persons were removed from the analysis, the DIF value was calculated. The effect of the existing DIF is determined by the magnitude of the DIF as indicated by the difference between two estimate values. The corresponding chi-square test was also obtained

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i2.34344

from DIF and standard errors of the estimates. If the chi-square is significant at a 0.01 level (Le, 2006) and the absolute DIF value is greater that 0.50 (Bond & Fox, 2015; Wu et al., 2016), an item is flagged for DIF. Based on the analysis, it is found that the chi-square (142.24, df = 49) is significant indicating the existence of DIF, and six of 50 items are flagged for DIF (items 7, 13, 20, 25, 45, and 46). Using the a magnitude value of more than 0.64 as moderate to large DIF magnitude proposed by Boone, Staver, and Yale (2014) , three items (items 7, 25, and 56) were detected to have moderate to large DIF showing the magnitude of 0.66, 2.16, and 0.97 respectively.

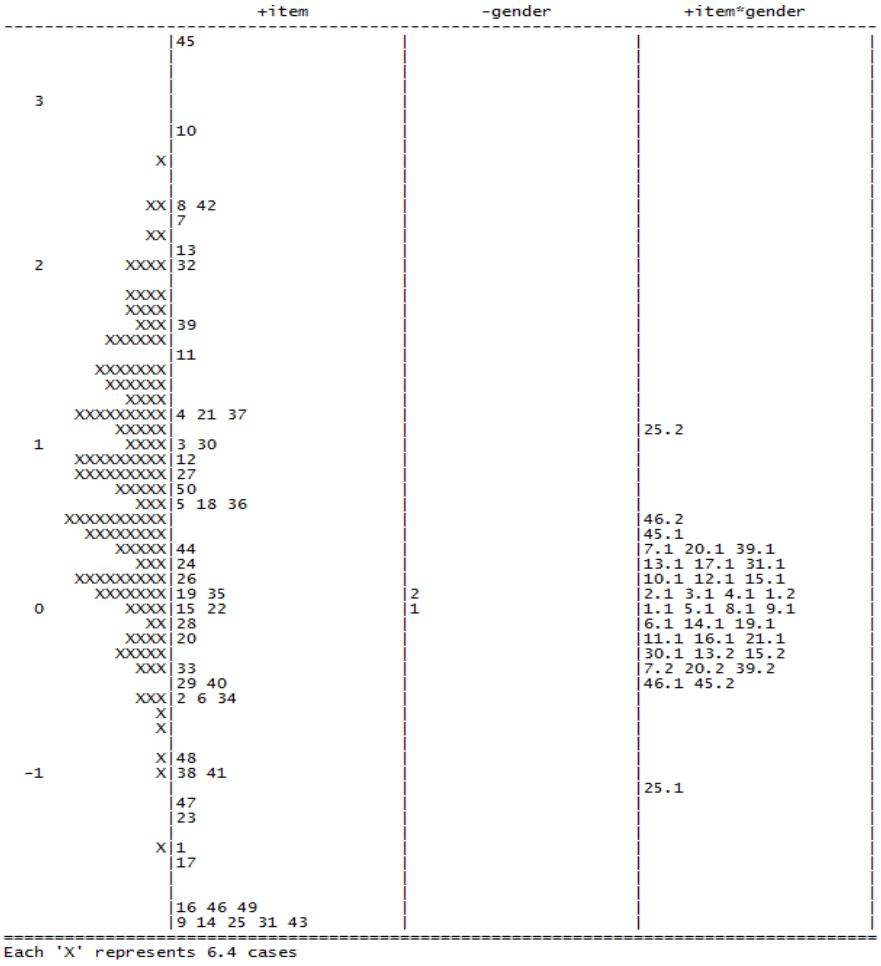

Figure 1. Wright Map of Item-Gender Interaction

Items 7, 25, and 46 are categorized into basic comprehension while items 13, 20, and 45 are vocabulary items. Of these six items, as Figure 1 shows, two items (items 25 and 46) display large magnitudes of 2.16 and 0.97 respectively which favor male students. Both items are in the category of basic comprehension items, particularly detailed and undetailed information. The reading passage of item 25 discusses a topic of entrepreneur and the correct response to the item indicates the information that is explicitly stated in the passage. Meanwhile, item 46 presents item that discusses foreign aid and the students are expected to answer an unstated detailed question. The potential source of this DIF may have been test takers unfamiliarity to the context of the passage. Table 3 presents the results of the gender DIF analysis by reading subskills. In terms of reading subskills, it is generally found that all test items slightly favored female more than male students. The main idea questions, for example, tended to be easier for male students than female students. This finding is not similar to the study of Pae (2012) presenting that the main idea items tended to be easier for female students.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i2.34344

Table 3. DIF Analysis by Reading Subskills

Reading Subskill	Item	Total Item	Easier for Males	Easier for Females
Reading for Main Idea (RMI)	1, 22, 33, 42	4	3	1
Reading for Basic comprehension (RBC)	2, 4, 6, 7, 9, 10, 11, 12, 16, 19, 21, 23, 25, 26, 29, 31, 32, 35, 39, 41, 46, 48, 49, 50	24	11	13
Inferencing (INF)	15, 17, 18, 28, 34, 40, 43,	7	3	4
Vocabulary Knowledge (VOC)	3, 5, 8, 13, 14, 20, 24, 27, 30, 36, 37, 38, 44, 45, 47,	15	7	8
	Total	**50**	**24**	**26**

To deal with the items detected for DIF, some approaches could be taken. Wu, Tam, and Jen (2016) suggest three possible ways of dealing with DIF items: removing the items, splitting DIF items for different groups, or leaving the items in the test. DIF items needed to be removed, as the further suggest, when many test items are available for selection into a final test. However, the consideration should be put on some influences to DIF, such as sample size. The more sample used to detect DIF, the higher the possibility to obtain higher DIF items. In addition to this, the items deemed to have DIF should be further examined to get substantive reasoning and make judgment about which items are actual DIF items. Le (2006) argues that items flagged for DIF are not necessarily deleted from future test, but the items need to be carefully reviewed. The DIF results provide information about the items that function differently in the test, and this information would be necessary for item writers to improve the item writing process (Zenisky, Hambleton, & Robin, 2003). For practical approach, Wu, Tam, and Jen (2016) suggest to use statistical analysis to identify DIF items and to examine the item content to investigate the theoretical explanations. In addition, the magnitude of the items should also be considered. They further argue that items with large DIF having more than 2 logits in item difficulty difference should be deleted.

CONCLUSION

Based on the DIF analysis of individual items in reading subtest in PTESOL, it is found that the subtest did not demonstrate much gender DIF. There were six out of fifty items in the subtest that were detected for gender DIF. Two items favored male students and four others favored female students. Reflecting from the findings, it is necessary to consider how the findings should be addressed in testing. Despite the fact that DIF is not equivalent to bias, the findings shown in the analysis would be indicative of potential sources of DIF, and the results can be utilized as guidelines for item writers to address what may be problematic in the item and what topics might be considered when developing test items.

REFERENCES

Adams, R., & Wu, M. (2010a). *ConQuest [Computer software]*. ACER.

Adams, R., & Wu, M. (2010b). *Differential Item Functioning*. ACER.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. S., & Yale, M. S. (2014). *Rasch Analysis in the human sciences*. Springer.

Curtis, D. D., & Boman, P. (2007). X-ray your data with Rasch. *International Education Journal*, *8*(2), 249–259.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v24i2.34344

Dodeen, H. (2003). The use of person-fit statistics to analyze placement tests. In *Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003)*.

Huff, K. L. (2000). *Evaluating Differential Item Functioning across selected item formats on a large-scale certification examination*. http://www.aicpa.org/BECOMEACPA/CPAEXAM/PSYCHOMETRICSANDSCORING/TECHNICALREPORTS/Pages/default.aspx

Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing*, *14*(3), 245–264. http://doi.org/10.1080/15305058.2013.877911

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298.

Kim, Y., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality, model-based DBF/DIF approach. *Language Learning*, *59*(4), 825–865.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly*, *24*(4), 741–746.

Kunnan, A. J. (2007). Test fairness, test bias, and DIF. *Language Assessment Quarterly*, *4*(2), 109–112. http://doi.org/10.1080/15434300701375865

Le, L. (2006). Analysis of Differential Item Functioning. In *The Annual Meetings of the American Educational Research Association in San Francisco, 7-11 April 2006*. Australian Council for Educational Research.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, *26*(2), 245–274.

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

Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring crossing differential item functioning by gender in mathematics assessment. *International Journal of Testing*, *15*(4), 337–355. http://doi.org/10.1080/15305058.2015.1057639

Pae, T. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*(4), 533–554. http://doi.org/10.1177/0265532211434027

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, *17*(3), 323–340.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*(1-2), 61-78. https://doi.org/10.1080/10627197.2004.9652959

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*(2), 136–147.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061