Tabaran Institute of Higher Education
International Journal of Language Testing
In press

ISSN 2476-5880
https://www.ijlt.ir

Investigating Gender DIF in the Reading Comprehension Section of the B2 First Exam

Chairil Anwar Korompot[1]*, Iskandarsyah Siregar[2] , Nurislom Iskandarovich Khursanov[3] , Diyorjon Abdullaev[4] , Khaled M. Mohamed[5]

ARTICLE INFO

Article History:
Received: October 2023
Accepted: December 2023

KEYWORDS
B2 First Exam
DIF
Mantel-Haenszel
Rasch model
Reading comprehension

ABSTRACT

Construct-irrelevant variance is considered as a major threat to validity which indicates the existence of additional unrelated variables that distort the meaning of test scores and cause the test to be biased. Differential item functioning (DIF) analysis is an important technique in examining the validity and fairness of educational tests. Concerning the importance of test fairness in large-scale exams, this study aimed to (1) detect gender DIF in the reading comprehension section of the B2 First exam using the Rasch model and Mantel-Haenszel method, and (2) investigate the comparability of results from the two DIF detection techniques. To this end, the reading section of the B2 First exam was administered to 207 undergraduate students of English as a foreign language (EFL). After checking the fit of the data to the Rasch model, the results of the Rasch model-based DIF analysis showed the presence of two items indicating DIF, whereas the results of Mantel-Haenszel showed that there were three gender-DIF items.

1. Introduction

Understanding written material in a second or foreign language (L2) is a highly intricate cognitive process, necessitating the synchronization of various cognitive and linguistic functions (Baghaei & Carstensen, 2013; Baghaei & Ravand, 2015; Boori et al., 2023; Elleman & Oslund, 2019). Proficiency in effectively comprehending written content in L2 is pivotal to the process of learning English. Therefore, evaluating reading skills holds great importance in diverse L2 programs and educational settings. Assessments of reading comprehension are extensively utilized in significant language proficiency examinations like the B2 First Exam and the International English Language Testing System (IELTS) globally, aiming to gauge the reading comprehension abilities of individuals aspiring to study or work in English-speaking environments. The scores from these assessments provide insight into the language proficiency of test takers and furnish credible data to guide decisions regarding each test taker. Given the impact that test scores' interpretations and decisions hold for all parties involved, developers and users of these tests prioritize the validity of the assessments.

In educational assessment and language evaluation, ensuring validity holds utmost significance in the test development process (AERA, 2014; Kane, 2013). For a test to be deemed effective in measuring individual attributes, it must demonstrate validity. A critical threat to validity is construct-irrelevant variance, as defined by Messick (1989). This term pertains to the variability in test-taker scores that can be attributed to irrelevant factors, distorting the intended meaning of the scores and thus diminishing the validity of the intended interpretation (AERA et al., 2014, p. 217). Essentially, a test taker's performance on a given test should remain unaffected by factors outside the scope of the intended

[1] Corresponding author: Universitas Negeri Makassar, Makassar, Indonesia, ORCID: 0000-0002-2006-906X, Email: ch.korompot@gmail.com
[2] Faculty of Language and Literature, Universitas Nasional, Indonesia, ORCID: 0000-0002-4529-6525
[3] Renaissance University of Education, Tashkent, Uzbekistan, ORCID: 0000-0001-5714-2745
[4] Department of Scientific Affairs, Innovations and Training of Scientific Pedagogical Personnel, Urganch State Pedagogical Institute, Urganch, Uzbekistan, ORCID: 0000-0001-8560-5604
[5] College of Mass Communications, Ajman University Ajman, 346, UAE, ORCID: 0000-0002-7194-942X

construct; otherwise, the test is considered biased (AERA, 2014). Differential Item Functioning (DIF) serves as a technique to identify such biases in a test.

As argued by Zumbo (2007), DIF occurs when the items of a test function differently for or against a particular group (e.g., gender, major, racial/ethnic, or nationality subgroups). In fact, an item is said to exhibit DIF if test takers with the same level of the expected construct have different probabilities of giving a correct response. The existence of DIF is an indication of multidimensionality which is considered a defect in the internal structure of a test (AERA, 2014). It must be noted that DIF is a prerequisite for bias. A biased item will definitely exhibit DIF; however, an item indicating DIF is not necessarily biased (Baghaei et al., 2017).

There are various statistical methods for detecting DIF such as logistic regression, Mantel-Haenszel, multiple-group factor analysis, and item response theory (IRT)/Rasch-based methods. Although all these methods intend to specify whether the test's functionality is tainted by an irrelevant factor, they do not function similarly in detecting items as DIF. Such diversity of methods for analyzing DIF might be bewildering for researchers and practitioners and might result in complexities of the findings derived from different DIF studies using various DIF detection techniques. Therefore, it is necessary to compare the results obtained from different DIF methods (Baghaei et al., 2019).

Taken together, the purpose of the present study is twofold. First, it aims to detect gender DIF in the reading comprehension section of the B2 First exam using the Rasch model (Rasch, 1960) and the Mantel-Haenszel method (Mantel & Haenszel, 1959), as the two most-commonly used DIF detection methods. Second, it aims to examine the comparability of results from the two DIF detection techniques.

2. Review of Literature

The analysis of Differential Item Functioning (DIF) in the context of reading comprehension has a well-established history, with numerous studies examining DIF in reading comprehension tests using observable factors such as age, gender, and native language. For instance, Pae (2011) employed various analytical approaches including linear multiple regression analysis, Mantel-Haenszel, and IRT to investigate gender-based DIF in reading comprehension. Their findings revealed several items displaying DIF due to the interactions between gender and item types. In a different study, Cadime et al. (2014) utilized logistic regression and Mantel-Haenszel DIF techniques to assess the influence of geographical region on DIF in a reading comprehension test taken by students from both urban and rural areas. While they identified 17 cases of non-substantive DIF out of 30, the overall integrity of the test results remained unaffected. Gnaldi and Bacci (2016) also employed a multidimensional latent class 2-PL (two-parameter logistic) IRT to detect DIF based on both gender and region. Their examination encompassed five latent classes across a comprehensive test battery covering grammar, reading, and mathematics, with the justification of latent classes incorporating covariates at both school and student levels. Another study by Geramipour (2019) utilized an item-focused trees approach to discern non-uniform and uniform DIF in an Iranian English as a foreign language (EFL) reading comprehension test, focusing on gender and academic background. Results from this study indicated that 10 items exhibited uniform DIF, with 2 items having 2 joint DIF predictor variables (2 splits), and 8 items having a single split. Conversely, non-uniform DIF analysis revealed 6 splits and 5 non-uniform DIF items, with only 1 item exhibiting 2 simultaneous DIF source variables. Additionally, Geramipour noted a significant correlation between background knowledge, gender, and EFL reading comprehension.

Amirian et al. (2020) analyzed DIF of the reading comprehension section of the Iranian National University Entrance Exam (INUEE) based on gender using Mantel-Haenszel. It turned out that only six items in the reading section flagged DIF, but their effect size was negligible. They also identified test takers' attitudes towards potential sources of DIF and unfairness. Results showed that test takers considered the test a fair one. Tabatabaee-Yazdi (2020) also applied the Hierarchical Diagnostic Classification Model (HDCM) to the reading comprehension section of the INUEE. The test was analyzed using HDCM and the Generalized Deterministic, Inputs, Noisy “and” Gate (GDINA; de la Torre, 2011) model to specify and compare test takers’ attribute mastery profiles in the test’s predefined skills and to indicate the associations among the attributes underlying the test to specify the sequence of teaching materials on increasing the probability of responding correctly to a set of test items. Moreover, Tabatabaee-Yazdi analyzed DIF to investigate whether the test functions equally across different subpopulations in terms of the test takers’ gender. It was found that one of the HDCMs and

GDINA fitted the data well. Yet, even though the analysis of HDCM revealed that there are dependencies among attributes of the reading comprehension test, the relative fit indices indicated a significant difference between the HDCM and GDINA. The analysis of DIF further showed a significant difference between females and males in six items; females outperformed males. She finally concluded that the IUEE test is a reliable and valid test.

Overall, very little attention has been paid to the analysis of DIF in tests of the University of Cambridge ESOL (English for Speakers of Other Languages) Examination Syndicate (e.g., Abbott, 2007; Aryadoust, 2012; Breland et al., 2004; Conoley, 2004; Geranpayeh & Kunnan, 2007; Ghaleb et al., 2023). Geranpayeh (2001) argued that more attention should be devoted to the analysis of gender-, nationality-, and age-related DIF in the ESOL tests. As stated by Geranpayeh and Kunnan (2007), tests constructed by the University of Cambridge ESOL Examination Syndicate are more likely to indicate DIF for these variables as gender, age, and nationality. They highlighted that "there has been a shift in the traditional test population, where test takers of many age [nationality and gender] groups are taking these cognitively challenging tests" (Geranpayeh & Kunnan, 2007, p. 193). In the absence of such studies, all stakeholders, researchers, and test users are unable to assume that the tests are fair and do not exhibit DIF for or against a particular group of test takers. To fill this research gap, this study aimed to apply the Rasch model and the Mantel-Haenszel method to analyze DIF of the reading comprehension section of the B2 First Exam based on gender.

3. Method

3.1. Participants and Setting

Data analyzed in the current study included item responses of 207 undergraduate students of EFL at the Faculty of Language and Literature, Universitas Nasional, Indonesia. There were 101 females (48.79%) and 106 (51.21%) males whose ages ranged from 19 to 33 (M = 22.45, SD = 3.89). Participation in the test was voluntary, and students' written consent was obtained for the study.

3.2. Instrumentation

A retired version of the reading comprehension section of the B2 First exam was used for the purpose of this study. The test totally consisted of 52 items which include multiple-choice cloze, open cloze, word formation, key word transformations, multiple-choice, gapped text, and multiple matching. Due to the different structures of items, sixteen items were removed from the analysis, and the remaining 36 items were analyzed.

4. Results and Discussion

4.1. Item Characteristics

Over the last three decades, numerous DIF detection methods have been proposed which vary from simple methods on the basis of difficulty indices (transformed item difficulty index or delta plot) to complicated techniques based on IRT models. For the purpose of this study, two methods were used: Mantel-Haenszel (Mantel & Haenszel, 1959) and the Rasch model (Rasch, 1960). Mantel-Haenszel is a nonparametric method for detecting DIF which is based on the concept of odds ratio. In other words, it is a chi-squared contingency table-based approach which analyzes differences between the reference and focal groups on all items of a test. The *Rasch model*, also known as the one-parameter logistic IRT model, is a psychometric technique that models the probability of a specified response (i.e., wrong/right answer) as a function of a test taker's ability and item difficulty (Aryadoust et al., 2021). A test taker with a greater ability has a higher probability to get a given item right. An important property of the Rasch model is measurement invariance. It states that the relationship between the ability level of test takers and their response to an item should be consistent across different groups and conditions. Put it differently, item difficulty and person ability parameters should be invariant across groups or populations, suggesting that the items are measuring the same latent trait in each group, and that differences in item responses are due to differences in ability levels of test takers, not their group membership. To check DIF of the reading comprehension section of the B2 First exam using the Rasch model and Mantel-Haenszel, the WINSTEPS software Version 3.73 (Linacre, 2009a) was employed.

Table 1 shows item difficulties, their standard error of measurement, fit indices, and point-measure correlation. Column two presents the difficulty of the test items, which indicates the position

of items on the construct continuum. As can be seen, item difficulties ranged from -2.54 to 2.63 logits, with item reliability and separation coefficients of 0.97 and 5.77, respectively. Moreover, person estimates ranged from -2.19 to 3.60, with person reliability and separation coefficients of 0.73 and 1.63, respectively. Item and person reliability coefficients demonstrate to what extent the test accurately measures item difficulties and person performance (Linacre, 2009b). Column three gives the error of measurement of item difficulties that indicate to what extent the estimation of item difficulties was accurate. Columns four and five show the results of infit and outfit mean squares (MNSQs), respectively. According to Linacre (2002), infit MNSQ is sensitive to inliers (e.g., anomalous behavior of items close to persons' measures), and outfit MNSQ is sensitive to outliers. The values of infit and outfit MNSQs indicated that four items (e.g., 4, 13, 22, and 25) did not fall within the ideal range of 0.50 and 1.50 (Bond & Fox, 2015; Linacre, 1999b). The last column depicts point-measure correlations for all the items. It shows to what degree observed scores are in agreement with the expected construct. Most of the values were positive and above 0.30, indicating the conformity of the items' difficulties and the Rasch model (Linacre, 2009b).

Table 1
Item Characteristics and Fit Statistics for the Reading Items

Items	Item Difficulty	Standard Error of Measurement	Infit MNSQ	Outfit MNSQ	Point-Measure Correlation
1	-0.38	0.19	1.32	1.33	0.04
2	0.10	0.17	1.22	1.31	0.10
3	1.45	0.15	1.02	1.30	0.25
4	2.05	0.15	1.16	1.73	0.05
5	2.63	0.17	1.03	1.36	0.18
6	-0.05	0.18	1.21	1.23	0.13
7	1.01	0.15	1.09	1.36	0.18
8	0.50	0.16	1.16	1.24	0.16
9	1.89	0.15	0.98	1.08	0.30
10	1.42	0.15	1.07	1.48	0.18
11	-0.70	0.21	1.31	1.46	0.03
12	-0.38	0.19	1.20	1.17	0.16
13	2.46	0.17	1.07	1.99	0.09
14	0.32	0.16	1.15	1.24	0.17
15	1.12	0.15	1.04	1.01	0.29
16	-1.56	0.28	0.77	0.85	0.48
17	-0.42	0.19	0.99	1.19	0.31
18	1.32	0.15	1.00	1.04	0.31
19	-0.42	0.19	0.95	1.04	0.37
20	0.35	0.16	0.94	0.89	0.42
21	0.32	0.16	0.97	0.99	0.38
22	-1.72	0.30	0.64	0.30	0.67
23	0.55	0.16	0.90	0.83	0.46
24	-1.10	0.24	0.87	0.75	0.48
25	-2.54	0.40	0.58	0.10	0.66
26	-0.01	0.17	0.96	0.96	0.40
27	-0.24	0.18	0.82	0.72	0.55
28	0.43	0.16	0.84	0.78	0.53
29	-0.34	0.19	0.97	1.11	0.34
30	0.02	0.17	0.92	0.86	0.45
31	-0.01	0.17	0.92	0.85	0.45
32	-1.21	0.25	0.73	0.56	0.61
33	-1.72	0.30	0.74	0.50	0.55

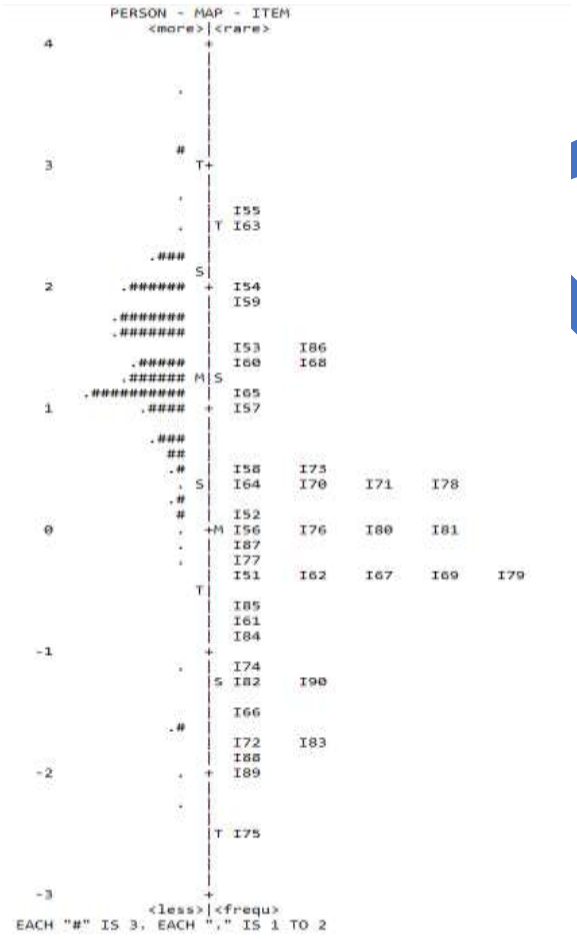
34	-0.84	0.22	0.79	0.76	0.55
35	-0.57	0.20	0.92	0.91	0.42
36	1.55	0.15	1.03	0.99	0.29

Note. MNSQ = Mean Square

The unidmensionality of the test was also investigated using the principal component analysis of linearized Rasch residuals (PCAR). Residuals are unexpected parts of the data and are differences between Rasch model expectations and the observed data. They are expected to be uncorrelated. The results of the PCAR analysis showed that the strength of the first contrast was 2.

Figure 1 illustrates the item-person map, known as the Wright map, indicating the relationship between item difficulties and test takers' performance along a single continuum. The map shows that the items of the test cover a wide range of difficulties, but they were easy for most of the test takers.

Figure 1
Item-Person Map



4.2. Comparison of the Rasch Model and the Mantel-Haenszel Method

Table 2 demonstrates the results of the Rasch model and Mantel-Haenszel method in detecting items exhibiting DIF. Columns labeled ‘Person Class’ show the class of test takers, e.g., males (M) and females (F). ‘DIF Measure’ columns present the difficulty of the test items
for the two classes. The more difficult the item, the higher the DIF measure is. As can be seen, Items 5 and 13 were the most difficult items in both male and female classes. ‘DIF Contrast’ gives the difference

in item difficulty between the two classes. This should be at least 0.5 logits for DIF to be noticeable (Linacre, 2009b). A positive DIF contrast shows that the item is more difficult for females, and a negative DIF value indicates that the item is more difficult for males. Tale 2 also shows a Welch *t* and a Welch probability. Welch *t* gives the DIF significance as a Welch's (Student's) *t*-statistic. As stated by Linacre (2009b, p. 424), "the *t*-test is a two-sided test for the difference between two means (i.e., the estimates) based on the standard error of the means (i.e., the standard error of the estimates)". The null hypothesis is that the two estimates are the same, except for measurement error. 'Prob.' indicates the probability of observing the amount of contrast by chance, when there is no systematic item bias effect. For statistically significant DIF on an item, a probability of ≤ 0.05 is required. Significant DIF values indicate that there is an irrelevant construct underlying the test.

As can be seen in Table 2, there were two significant gender-based DIF items (e.g., Items 10 and 11). The difficulty of Item 10 is 0.87 for females and 1.94 for male test takers. The contrast is also significant (DIF Contrast = -1.07 > 0.50, and *p*-value = 0.00 < 0.05). Similarly, the difficulty of Item 11 is -1.25 for females and -0.30 for male test takers. The contrast is significant (DIF Contrast = -0.95 > 0.50, and *p*-value = 0.032 < 0.05).

Finally, the last two columns of Table 2 illustrate the results of Mantel-Haenszel. As one can see, three items (e.g., Items 10, 20, and 35) had significant values, indicating the presence of DIF.

Table 2
Comparison of Rasch Model and Mantel-Haenszel Method

	Rasch Model							Mantel-Haenszel	
Items	**Person Class**	**DIF Measure**	**Person Class**	**DIF Measure**	**DIF Contrast**	**Welch t (d.f.)**	**Prob.**	**Chi-square**	**Prob.**
1	F	-0.18	M	-0.60	0.42	1.09 (204)	0.276	3.725	0.054
2	F	-0.11	M	0.28	-0.39	-1.13 (203)	0.258	0.003	0.957
3	F	1.51	M	1.38	0.13	0.44 (204)	0.663	0.708	0.400
4	F	2.13	M	1.98	0.15	0.48 (204)	0.633	0.554	0.457
5	F	2.89	M	2.42	0.47	1.35 (203)	0.179	1.397	0.237
6	F	0.02	M	-0.11	0.13	0.37 (204)	0.713	0.469	0.493
7	F	0.82	M	1.18	-0.35	-1.18 (204)	0.238	0.010	0.921
8	F	0.53	M	0.48	0.05	0.15 (204)	0.877	0.116	0.733
9	F	2.13	M	1.68	0.45	1.49 (204)	0.139	1.635	0.201
10	F	0.87	M	1.94	-1.07	3.52 (204)	0.000	8.851	0.003
11	F	-1.25	M	-0.30	-0.95	-2.16 (198)	0.032	1.480	0.224
12	F	-0.32	M	-0.44	0.13	0.33 (204)	0.740	0.999	0.317
13	F	2.51	M	2.42	0.09	0.27 (204)	0.785	0.664	0.415
14	F	0.42	M	0.23	0.19	0.59 (204)	0.554	0.000	0.996

15	F	0.92	M	1.30	-0.38	-1.29 (204)	0.198	1.070	0.301
16	F	-1.13	M	-2.25	1.12	1.78 (197)	0.077	3.072	0.080
17	F	-0.63	M	-0.24	-0.40	-1.02 (203)	0.309	0.877	0.349
18	F	1.15	M	1.47	-0.32	-1.08 (204)	0.283	0.933	0.334
19	F	-0.24	M	-0.60	0.35	0.91 (204)	0.366	0.014	0.906
20	F	0.14	M	0.52	-0.38	-1.18 (203)	0.241	4.948	0.026
21	F	0.26	M	0.38	-0.12	-0.38 (204)	0.704	0.000	0.989
22	F	-1.38	M	-2.25	0.87	1.35 (200)	0.179	0.012	0.913
23	F	0.37	M	0.71	-0.34	-1.08 (203)	0.281	0.872	0.350
24	F	-0.91	M	-1.31	0.40	0.82 (204)	0.413	0.015	0.901
25	F	-2.04	M	-3.74	1.70	1.52 (181)	0.129	-	-
26	F	0.02	M	-0.05	0.07	0.20 (204)	0.843	0.011	0.917
27	F	-0.11	M	-0.37	0.26	0.71 (204)	0.478	0.119	0.730
28	F	0.47	M	0.38	0.09	0.29 (204)	0.769	0.005	0.941
29	F	-0.11	M	-0.60	0.49	1.28 (204)	0.203	1.624	0.202
30	F	0.08	M	-0.05	0.13	0.38 (204)	0.707	0.854	0.355
31	F	-0.04	M	0.01	-0.05	-0.15 (204)	0.880	0.180	0.672
32	F	-1.13	M	-1.31	0.18	0.36 (204)	0.717	0.045	0.831
33	F	-1.52	M	-2.00	0.48	0.78 (203)	0.438	0.033	0.855
34	F	-0.63	M	-1.07	0.43	0.98 (204)	0.329	0.010	0.921
35	F	-0.91	M	-0.30	-0.61	-1.49 (201)	0.139	3.997	0.046
36	F	1.60	M	1.51	0.10	0.32 (204)	0.747	0.005	0.944

Note. F = Female; M = Male; DIF S.E. = Standard error of DIF; d.f. = Degrees of freedom; prob. = Probability

5. Conclusion

This study set out to both use the Rasch model and the Mantel-Haenszel method to investigate gender DIF in the reading comprehension section of the B2 First exam and investigate the comparability of results from the two DIF detection techniques. As the analysis of the Rasch model showed, there were four misfitting items (e.g., 4, 13, 22, and 25). The Wright map also indicated that most of the test items were easy for the test takers, so more difficult items should be added to the test to align with the ability level of all test takers.

After checking the fit of the data to the Rasch model, the results of Rasch model-based DIF analysis showed the presence of two items (e.g., Items 10 and 11) exhibiting significant DIF. The two items were easier for female test takers than the male ones. On the other hand, the results of Mantel-Haenszel showed that there were three significant gender-DIF items (e.g., Items 10, 20, and 35), indicating that the items were more difficult for male test takers compared to female ones. Based on the findings of the study, the results of DIF analysis using the Rasch model and Mantel-Haenszel method show that although a small number of items were biased, a very large part of the present test was unbiased against the test takers, and it can be concluded that the test is fair. It is worth mentioning that the concept of fairness is multidimensional, and it can be affected by different factors which should be controlled (Aryadoust, 2012). The results of this study converge with earlier studies (e.g., Amirian et al., 2020; Geramipour, 2019; Gnaldi & Bacci, 2016; Pae, 2011) in which gender was exhibited as a viable bias source in reading tests. The analysis of items exhibiting DIF across the two methods showed that the items assess test takers' ability to make paraphrases. Low-ability male test takers were also attracted to some item distractors and tended to guess the correct answer on the test items. Research on cognitive psychology has indicated that males have a greater tendency to take more risks when they face a problem compared to females (Richardson, 1997).

Furthermore, the results of this study generally show that the two methods produce different patterns in detecting DIF. This reflects that the two methods differ in their underlying assumptions, methods of analysis, and the types of DIF they are designed to identify. The choice between the two methods often depends on the specific context of the study and the characteristics of the data. In the context of DIF analysis using the Rasch model, the focus is on item bias, and the analysis aims to identify whether items function differently for different groups of individuals after accounting for their overall ability levels. The Rasch model requires relatively large sample sizes, and it assumes that the latent trait (ability) is measured on an interval scale. The model also assumes measurement invariance, meaning that the relationship between item responses and the underlying trait (ability) is consistent across different groups. On the other hand, the Mantel-Haenszel method is more general and can be applied to detect DIF in both items and tests. It examines the association between an item response and group membership while controlling for a third variable (usually an ability variable). The method is less restrictive in terms of sample size requirements and assumptions about the measurement scale, and does not assume measurement invariance. It simply assesses whether the association between item responses and group membership is consistent across levels of a third variable. It must also be noted that the Mantel-Haenszel method is assumed to be more efficient for very easy or very difficult items, while its efficiency is reduced for tests with moderate difficulty.

When considering the results from this study, limitations and suggestions for future research are important to note. The sample of the present study (N=207) was not very impressive for detecting DIF in the B2 First exam as a standardized large-scale test. Future studies can adopt larger sample sizes. The present study extends DIF literature by offering information on DIF with an Indonesian sample. The results of DIF analyses might be affected by nationalities. It is thus recommended for future studies to examine gender DIF among different nationalities. Moreover, future studies can use more DIF detection techniques to confirm the research findings of the current study and explore any undetected DIF items due to the shortcomings of each of the utilized methods. It is also important to conduct further studies to check whether these results can be substantiated with a larger pool of test items. More importantly, items exhibiting DIF need to be meticulously analyzed and modified. Future studies can carry out a content analysis to identify the main sources of DIF. Finally, further studies can be carried out in various test situations to allow test developers and researchers to improve or remove biased items.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no specific funding for this work from any funding agencies.

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*1(1), 7–36. https://doi.org/10.1177/0265532207071510

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. AERA.

Amirian, S. M. R., Ghonsooly, B., & Amirian, S. K. (2020). Investigating fairness of reading comprehension section of INUEE: Learner's attitudes towards DIF sources. *International Journal of Language Testing*, *10*(2), 88–100. URL:https://www.ijlt.ir/article_118016.html

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of International English Language Testing System (IELTS) listening module. *International Journal of Listening*, *26*(1), 40–60. https://doi.org/10.1080/10904018.2012.639649

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, *18*, 1–13. Available at:https://scholarworks.umass.edu/pare/vol18/iss1/5

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences, 43,* 100–105. https://doi.org/10.1016/j.lindif.2015.09.001

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology*, *19*, 155–168. URL:http://najp.us/north-american-journal-of-psychology-index

Baghaei, P., Ravand, H., & Nadri, M. (2019). Is the d2 test of attention Rasch scalable? Analysis with the Rasch Poisson Counts Model. *Perceptual and Motor Skills*, *126*(1), 70–86. https://doi.org/10.1177/0031512518812183

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

Boori, A., Ghazanfari, M., Ghonsooly, B., & Baghaei, P. (2023). The construction and validation of a Q-matrix for cognitive diagnostic analysis: The case of the reading comprehension section of the IAUEPT. *International Journal of Language Testing*, *13*(Special Issue), 31–53. https://doi.org/10.22034/ijlt.2023.383112.1227

Breland, H., Lee, Y.-W., Najarian, M., & Muraki, E. (2004). *An analysis of the TOEFL CBT writing prompt difficulty and comparability of different gender groups* (ETS Research Rep. No. 76). Educational Testing Service.

Cadime, I., Viana, F. L., & Ribeiro, I. (2014). Invariance on a reading comprehension test in European Portuguese: A differential item functioning analysis between students from rural and urban areas. *European Journal of Developmental Psychology*, *11*(6), 754–766. https://doi.org/10.1080/17405629.2014.938629

Conoley, C. A. (2004). *Differential item functioning in the Peabody Picture Vocabulary Test-Third Edition: Partial correlation versus expert judgment*. Texas A&M University.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

Elleman, A. M., & Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences*, *6*(1), 3–11. https://doi.org/10.1177/2372732218816339

Geranpayeh, A. (2001). *Country bias in FCE listening comprehension* (Cambridge ESOL Internal Research and Validation Report No. 271). University of Cambridge.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), 190–222. https://doi.org/10.1080/15434300701375758

Gnaldi, M., & Bacci, S. (2016). Joint assessment of the latent trait dimensionality and observed differential item functioning of students' national tests. *Quality & Quantity*, *50*(4), 1429–1447. https://doi.org/10.1007/s11135-015-0214-0

Ghaleb, M., Suleiman, O., Mohammed, A., Alqiraishi, Z. H. A., Mutar, H. K., Mohammed Ali, Y., Hadi, F., Anber, A. A., Emaimo, J., & Karandeeva, L. G. K. (2023). Evaluating measurement invariance in the IELTS listening comprehension test. *International Journal of Language Testing*, *13*(Special Issue), 134–141. https://doi.org/10.22034/ijlt.2023.399416.1254

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–173. https://doi.org/10.1111/jedm.12000

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103–122. URL: http://jampress.org/jom.htm

Linacre, J. M. (2009a). Winsteps® Rasch measurement computer program (Version 3.73). Portland, Oregon: Winsteps.com.

Linacre, J. M. (2009b). *A user's guide to winsteps*. Chicago, IL: Winsteps.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed., pp. 13–103). American Council on Education and National Council on Measurement in Education.

Pae, H. (2011). *Differential item functioning and unidimensionality in the Pearson Test of English Academic*. Pearson Education Ltd. Retrieved from https://pearsonpte.com/wp-content/uploads/2014/07/RN_Differential-ItemFunctioning.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Danish Institute for Educational Research.

Richardson, J. T. E. (1997). Gender differences in cognition: Results from meta-analysis. In P. A. Caplan, M. Crawford, J. S. Hyde, & J. T. E. Richardson (Eds.), *Gender differences in human cognition* (pp. 3–29). Oxford University Press.

Tabatabaee-Yazdi, M. (2020). Hierarchical diagnostic classification modeling of reading comprehension. *SAGE Open*, *10*(2). 1–13. https://doi.org/10.1177/2158244020931068

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.org/10.1080/15434300701375832