Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 10, No. 2, October 2020

ISSN 2476-5880

Investigating Fairness of Reading Comprehension Section of INUEE: Learner's Attitudes towards DIF Sources

Seyed Mohammad Reza Amirian[1*],
Behzad Ghonsooly[2], Seyedeh Khadijeh Amirian[3]

Received: 12 June 2020 Accepted: 13 September 2020

Abstract

The purpose of the present study was two-fold: (a) First, it examined fairness of Special English Test (SET) of Iranian National University Entrance Exam (INUEE) by analyzing Differential Item Functioning (DIF) with reading comprehension section of this test (b) second, it explored test takers' attitudes towards possible sources of unfairness and DIF. In the quantitative part of the study the data from 10000 test takers (6820 females and 3180 males) were analyzed for gender DIF using Mantel-Haenszel (MH) technique. It was revealed that only 6 items in the reading comprehension skill showed DIF. Further analysis manifested that the effect size of DIF for all six items were category A or negligible. Moreover, qualitative interview results indicated that learners generally considered the test a fair one while some potential sources of bias such as *topic familiarity, multiple-choice format of the test, topic interest, passage length, and complex structure of test items* were mentioned.

Keywords: Assessment**,** DIF, Fairness, Validity, DIF Sources

1. Introduction

Test fairness in language assessment has been now discussed by researchers for over three decades. Fairness is characterized by the absence of bias towards any identifiable group of test takers. Test fairness is a central concern in psychological testing which has to do with procedures for selecting, administering, and interpreting test scores in an applied setting. Some researchers, for instance Xi (2010) and Kunnan (2010), treat fairness as an aspect of validity and conceptualize it as comparable validity for different groups.

The issue of test fairness has attracted a lot of attention recently because it is directly related to the validity of the test and test takers with different backgrounds from across the world expect to be fairly treated by language tests. Therefore, many testing organizations such

[1*] Associate Professor of Applied Linguistics; Department of English Language and Literature, Hakim Sabzevari University. Email: sm.amirian@hsu.ac.ir
[2] Professor of Applied Linguistics; Department of English Language and Literature, Ferdowsi University of Mashhad, Mashhad.
[3] MA in Educational Psychology; Department of Psychology, Ferdowsi University of Mashhad.

as Educational Testing Service (ETS) have published standards for equity and fairness (ETS, 2002) to ensure the tests are fair to all participating test takers.

In psychometric bias analysis, Differential Item Functioning (DIF) analysis has become the new standard (Zumbo, 1999). DIF analysis is commonly conducted in a large-scale assessment and a standardized test to ensure that all test takers are treated fairly by the test regardless of their group membership. DIF items pose a serious threat to the validity of a test, that is why DIF detection has become an important stage in test validation process (Zumbo, 2007) especially in standardized testing contexts (Pae & Park, 2006).

DIF exists if individuals in the focal group (i.e., males in this study) find an item more difficult than the individuals in the reference group (i.e., females) after ability and other factors have been controlled. In other words, the items function differently within these groups (Camilli & Shepard, 1994); therefore, it seems necessary to apply some analytical methods for detecting and removing the items flagged with DIF. Mantel-Haenszel (MH) procedure was first proposed by Mantel and Haenszel (1959) in their seminal paper. Holland and Thayer (1988) adapted this method for detecting differential item functioning. MH is a nonparametric approach for identifying DIF which belongs to contingency table related approaches. It is based on contingency tables and observed conditioning variable which explicitly matches the examinees from two different groups on the ability of interest, and then compares the likelihood of success on the item for the two groups.

Numerous studies have investigated differential performance of items on high-stakes standardized tests such as TOEFL (Hill & Liu, 2012; Lee, Brenald, & Murki, 2004) and IELTS (Aryadoust, 2012) to make sure test items are not differentially favoring a particular group of test takers. Within Iranian context, also, DIF analysis has become a common procedure in designing tests such as University of Tehran English Proficiency Test (UTEPT) (Alavi, Rezaee, & Amirian, 2011; Amirian, Alavi, & Fidalgo, 2014; Fidalgo, Alavi, & Amirian, 2014), and Iranian National University Entrance Exam (INUEE) (Barati, Ketabi, & Ahamdi, 2006, Geramipour, 2020).

INUEE is a national high-stakes test in Iran taken by more than one million examinees every year making it the most influential test in Iran affecting large number of high school students hoping to grant a seat at Iranian universities. Now, the presence of DIF within such a vital examination can liquidate the validity of INUEE. To ensure all the stake holders including families, test takers and administrators about the fairness of this national high-stakes test, it is necessary to conduct a precise and detailed gender DIF analysis of the test.

The focus of this study is on the reading comprehension skill of Special English Test (SET) of INUEE because previous research shows that learners' background such as gender, academic background, etc. play a pivotal role in reading comprehension (Amirian, Alavi, & Fidalgo, 2014; Hill & Liu, 2012; Pae, 2012). Toker (2019) argues that 'topic effect' or 'content knowledge' might affect the reading scores in the TOEFL iBT which threatens the validity of the entire reading skill. Moreover, content knowledge is a critical aspect of reading comprehension ability which can be a potential source of bias (Li, Liu, & Steckelberg, 2009; Keshavarz, Atai, & Ahmadi, 2007).

Gender DIF is a major source of bias that can threaten the validity of a standardized test such as INUEE and render it biased. Although gender DIF as one of the most important group difference factors has been extensively researched in DIF literature both in Iran (Abdorahimzadeh, 2014; Amirian, Alavi, & Fidalgo, 2014; Barati, Ketabi, & Ahamdi, 2006; Geramipour & Shahmirzadi, 2019) and overseas (Pae, 2012; Aryadoust, Goh, & Kim, 2011; Park & French, 2013; Pae, 2004, 2012), few studies have focused on DIF analysis of the reading comprehension section of INUEE to detect items performing differentially for males or females. To fill this gap, the first purpose of the current study is to analyze items of INUEE for potential DIF.

Finding DIF items is not sufficient for detecting bias and further analysis of DIF flagged items is required to uncover the underlying reasons for differential performance of items. Although many studies have addressed gender DIF (e.g. Barati, Ketabi & Ahmadi; Kunnan, 1990), very few studies have embarked on discovering the potential causes of DIF (Jalili, Barati, & Moein Zadeh, 2020; Lee & Geisinger, 2014; Suh & Talley, 2015; Wue & Erickan, 2006) because it's sometimes really hard to find causes of DIF in the test item itself or other contextual variables which are at play. In other words, it must be determined if sources of DIF is a construct-relevant or construct-irrelevant factor to distinguish *bias* from *impact* and to inform future test development (Zenisky, Hambleton, & Robin, 2004).

Underlying sources of DIF could be attributed to a plethora of variables. Zumbo et al (2015) used the term *ecology* of item responding to refer to all influential factors including

> (a) test format, item content, and psychometric dimensionality; (b) person characteristics and typical individual differences variables such as cognition; (c) teacher, classroom, and school context; (d) the family and ecology outside of the school; and finally (e) characteristics of the community, neighborhood, state, and nation. (p. 139)

Nonetheless, learners' views are not accounted for in this model which is a gap in the literature on DIF that is addressed by the present study.

Multiple choice format is the only format for all questions on INUEE including reading comprehension skill. In terms of underlying causes of DIF in reading comprehension items based on question format some ideas are put forward (Jalali, Barati & Moein Zadeh, 2020; Pae, 2012; Taylor & Lee, 2012). For example, Taylor and Lee (2012) found that multiple-choice reading comprehension items generally favored males while constructed-response items generally favored females. Moreover, in flagged reading items which typically assessed text interpretations or implied meanings, males were shown to perform better on items that asked them to identify reasonable interpretations and analyses of informational text. However, the results on sources of DIF for reading items are inconclusive and the present study tends to discover the underlying reasons for DIF from learner's perspective.

Considering the paucity of research on underlying reasons of DIF in general and causes of DIF in reading comprehension items in particular, the purpose of the present study is to (a) detect DIF in reading comprehension section of Special English Test (SET) of INUEE and (b) to elicit test takers' attitudes towards the underlying reasons of unfairness in reading multiple-choice items. Although Zumbo et al (2015) in their model of ecology of test performance capture multiple factors that may cause DIF, they fail to address the issue from learners' perspectives to discover how test takers see the test items and what sources of bias they experience in taking reading tests. Thus, this study intends to shed light on examinee's views towards sources of DIF in INUEE through interviews.

2. Literature Review

In the evaluation of tests, from the late 1990s test fairness is seen as a fundamental concept in the field of language assessment (Kunnan, 2010). Many assessment scholars perceive fairness as an overarching test quality that encompasses validity, accessibility to the test, absence of bias, social consequences, and conditions of administration (Banerjee, 2016).

A fair assessment happens when students are given equitable chance to show what they know (Lam, 1995). This does necessarily imply all students should be treated exactly the same, rather they should be assessed using methods and procedures which are most appropriate to them (Suskie, 2000). DIF happens if examinees from different groups do not have equal chance of excelling on the item after they are matched on the underlying knowledge that the item is measuring (Zumbo, 1999).

Scott, Webber, Lupart, Aitken, and Scott (2014) identified some major principles for fair assessment for all students. First, they indicated that teachers and professors must address the personal impact of assessment practices. Then, they explained that assessment practices should be differentiated in order to accommodate various social and cultural factors. Finally, they argued that frequency and intensity of tests must not be overwhelming for students.

Reading comprehension is one of the skills often analyzed for DIF in terms of content knowledge. In a study on TOEFL iBT, Liu, Schedl, Malloy, and Kong (2009) investigated the effect of content knowledge on reading performance. The results revealed that the majority of items displayed little or no DIF. Moreover, DIF was not observed in the passages of the study.

However, in a study with 240 Iranian male students who learned English as a foreign language, Keshavarz, Atai, and Ahmadi (2007) reported a significant relationship between familiarity with content and reading comprehension test scores and recall scores. Also, Geramipour (2020) reported significant relationships between gender and background knowledge with EFL reading comprehension. As the results of studies on contribution of content knowledge to reading performance have produced controversial results, in the present study, it is intended to examine whether reading section of INUEE show any differential performance for different gender groups, i.e. the content of reading passages systematically favors one gender over the other.

Moreover, studies on the causes of gender DIF in reading comprehension items are meagre in the literature. Pae (2012) analyzed gender DIF on the English subtest of the Korean College Scholastic Aptitude Test (KCSAT) over a nine-year period using the MH and item response theory likelihood ratio (IRT-LR) procedures. Reading strategy and perceived interest were reported as two factors that explained gender DIF. The results revealed an interaction between item type and gender DIF. Moreover, a significant relationship was reported between gender differences in the examinee's perceived interest in test items and the size of gender DIF.

In a meta-analysis, Koo, Becker, and Kim (2014), investigated reading test of Florida Comprehensive Achievement Test (FCAT) for third and tenth graders using MH DIF indices. They attempted to examine DIF trends for English language learners (ELLs) versus non-ELL students in third and tenth grades on a large-scale reading assessment. The results revealed that items requiring knowledge of words and phrases in context favored non-ELLs in grade 3, whereas items requiring evaluation skills favored ELLs in grade 10. Nonetheless, as for gender DIF, inconsistent patterns were found which signals the need for further research on gender DIF.

In a recent study, within Iranian EFL context, Jalili, Barati, and Moein Zadeh (2020) investigated DIF by gender grouping through logistic regression modeling in the TOEFL reading paper. Their results manifested three ecological variables as potential causes of DIF including income, administration convenience, and SES.

Although some of the reviewed studies in this part have made efforts to explain the underlying causes of DIF, there is a paucity of research on the underlying causes of DIF from test takers' perspective through interviews. In a rare study, Flores, Simao, Barros, and Pereira (2014) investigated the perceptions of students on the fairness of assessment methods concluding examinees considered those methods of assessment fairer which required their active involvement. Thus, considering the gap in literature on examinee's perceptions of fairness, the aim of the present research is to, first, identify DIF items in reading comprehension of SET of INUEE and, second, to elicit examinee's attitudes on test fairness and possible causes of DIF. The following questions are posed to be answered.

RQ1. Does reading comprehension section of Special English Test of INUEE show significant gender DIF?

RQ2. What are test takers' attitudes towards fairness of reading comprehension section and possible causes of gender DIF?

3. Method

3.1. Participants and Data Source

The data set for the quantitative phase of the study came from a sample of 10000 test takers who took the Iranian National University Entrance Examination (INUEE). The sample was divided into a focal group of 3180 males (31.8%), and a reference group of 6820 females (68.2%). Before running MH DIF, examinees were matched in accordance with their total test scores. Moreover, for the qualitative part of the study, 15 participants who had taken INUEE recently were selected through criterion sampling for semistructured interviews.

3.2. Procedure

This study was part of a large-scale study consisting of quantitative and qualitative phases. In the quantitative phase of the study, the MH method was used to detect differentially functioning items of the SET of INUEE. The MH DIF statistic is the most common measure for DIF. Moreover, the MH DIF effect size guidelines for measuring the size of DIF are known to researchers and practitioners. The data for the current study were of dichotomous nature. The

1-0 item responses for the test are from a sample of 10000 examinees who sat for SET part of INUEE. They were divided into a reference group of 6820 females and a focal group of 3180 males. The data for 20 multiple choice reading comprehension items were generated. After matching examinees on their ability level, the items were analyzed for DIF. Examinees' total score on INUEE was used as the matching criterion. The magnitude of DIF in each item was estimated by calculating the MH statistics. DIFAS program (Penfield, 2009) was employed to run MH DIF.

In the second phase of the study, to discover underlying causes of DIF, semistructured interviews were conducted with 15 freshmen English literature students (8 females and 7 males) who had taken the reading comprehension section of INUEE to grant a seat at public universities. To refresh their minds, the reading comprehension questions were presented to them again and their attitudes towards the fairness of reading items and possible sources of bias against a particular gender group were elicited. The following guiding questions were asked in interview sessions.

What are the characteristics of a fair reading comprehension test? Is reading section of SET of INUEE fair?

Do you think the reading comprehension items of SET of INUEE favor a particular gender group? Why?

Interviews were conducted in the mother tongue of the learners (Persian), to make sure they understand the questions and can respond freely. Afterwards, the recorded interviews were transcribed and then translated into English for further qualitative content analysis. Common themes and categories in responses were coded and results were reported.

4. Results and Discussion

4.1. MH DIF Results

The purpose of the quantitative part of this study was to analyze reading comprehension items of SET section of INUEE for gender DIF through MH procedure. The data set came from 10000 male and female test takers. DIF statistics results for the reading comprehension section of INUEE is summarized in Table 1. Negative values show the direction of DIF toward males (focal group) while the positive values show the direction of DIF toward females (reference group).

Table 1.
DIF Statistics for Reading Comprehension of INUEE

ITEM No.	MH CHI	MH LOR	BD	CDR	ETS
Item 1	0.21	-0.03	0	OK	A
Item 2	18.19	-0.31	0.29	Flag	A
Item 3	31.98	-0.37	0.18	Flag	A
Item 4	0.08	0.02	0.21	OK	A
Item 5	0.45	-0.05	4.34	OK	A
Item 6	3.10	-0.16	2.28	OK	A
Item 7	0.20	0.04	1.57	OK	A
Item 8	15.50	0.31	0.03	Flag	A
Item 9	2.09	0.15	0.42	OK	A
Item10	3.49	0.18	1.96	OK	A
Item11	28.81	-0.34	4.36	Flag	A
Item12	3.68	0.23	1.15	OK	A
Item13	11.16	0.20	0.03	Flag	A
Item14	3.56	0.21	0.98	OK	A
Item15	17.24	0.30	0.11	Flag	A
Item16	1.73	0.10	3.22	OK	A
Item17	0.41	-0.04	2.17	OK	A
Item18	2.52	0.13	1.57	OK	A
Item19	0.82	-0.08	0.33	OK	A
Item20	1.12	-0.08	0.30	OK	A

Note: A= negligible effect size

As indicated in the table, six items (30%) out of 20 in the reading comprehension section were flagged with DIF while 14 items were OK. Three items were in favor of females and three items in favor of males. As such, items 8, 13, and 15 were found to be functioning differentially for females, while items 2, 3, and 11 functioned to the advantage of males.

Thus, as Zumbo (2003) maintains DIF at the level of individual items may be canceled at the test level if the same number of items indicate DIF in favor of different groups. However, regardless of the direction of DIF, the DIF detected at the item level may lead to test bias. In other words, we may not delete DIF at the test level (Pae and Park, 2006), and hence need to carefully take into account all the items that indicate DIF.

Beside the number of items showing DIF, DIF cancellation depends on the magnitude of DIF (Pae, & Park 2006). In the current study, as the number of gender DIF items is regarded, of a total of 20 items DIFAS detected only three items favoring males and three items favoring females; hence, DIF cancellation might be at work at least at the level of number of DIF items favoring each group. Moreover, it was found that the magnitude or effect size of DIF does not indicate the existence of a significant level of DIF sacrificing the fairness of the reading skill in general (all six items were flagged as negligible or level A DIF).

Therefore, in terms of the size of gender DIF, it was found that all the DIF flagged items in this section were classified as the small size of DIF (category A). Also, the degree and number of DIF items in the reading comprehension section confirmed the absence of bias towards any gender group of test takers. This finding is in line with Barati, Ketabi, and Ahmadi (2006) who found similar number of items in INUEE favoring different groups of test takers in reading comprehension section.

In a DIF study similar to this one, Geramipour (2020) analyzed reading comprehension section of INUEE on a sample of 4937 MA candidates who took the test in 2015. He employed item-focused trees (IFT) to flag both uniform and non-uniform DIF items. His results indicated that 10 items displayed uniform DIF and five items displayed non-uniform DIF. The conclusion was that gender and background knowledge had significant relationships with EFL reading comprehension. However, in the present study gender was not found to be a significant irrelevant factor affecting learners' performance on reading comprehension section of INUEE, so the fairness of the reading section is not endangered. However, to cast further light on the causes of DIF in these six flagged items, examinee's perception of sources of DIF in these items were elicited.

4.2. DIF Sources Results

To achieve the second aim of the study that was uncovering the underlying causes of DIF in reading items of INUEE, qualitative content analysis was undertaken on 15 test- taker interviews. Test takers were asked to express their attitudes towards the fairness of all reading items and the possible underlying causes of differential performance of the six items that were flagged as showing DIF in the quantitative phase of the study. Overall, in line with the quantitative findings of the study, the test takers believed that the reading passages were fair to both genders which is consistent with the findings of Barati and Ahmadi (2010) who concluded that the reading comprehension section of INUEE favored males and females equally. However, they claimed that females were favored on grammar, language function, and the cloze test sections while males were favored on the vocabulary and word order sections.

A common theme that emerged in content analysis was that learners considered a fair reading test one that assessed their true language ability and could distinguish fairly high ability learners from low-ability ones. The transcriptions of some students' interviews are presented below (E stands for Excerpt)

E1 "I think, the reading section was fair because it tested my reading skill well. If a test measures my real abilities not what I have memorized, it is fair"

Another interviewee pointed out to the topic of reading comprehension passages as a major factor in the fairness of items. She believed that topics which are interesting are fair. This is consistent with Toker (2019) who argued that 'topic effect' influences the reading scores in the TOEFL iBT which threatens the validity of the entire reading skill.

E2 "I think the topic of reading comprehension passages should be interesting for us because sometimes I am bored when I read a passage."

Two interviewees referred to the topic of reading passages as well but they pointed to topic familiarity as a source of bias. This finding is consistent with Geramipour (2020) who stressed examinees' field of study is a background variable that is directly related to the background knowledge and is consequently associated with EFL reading comprehension. Amirian, Alavi, and Fidalgo (2014) also found test takers' field of study a source of DIF with humanities-oriented subjects rated as favoring females and science-oriented subjects rated as favoring males.

E3 "Sometimes the reading passages come from unfamiliar fields such as philosophy, geography, medicine, etc. which are irrelevant to our fields of study and difficult to understand."

E4 "The first reading comprehension which came from astrology was very difficult to me as a female test taker. I suppose this passage favors males because they are better at these subjects."

Another potential source of differential performance for test takers was the complex structure of items. For example, item 2 as a reading item flagged with DIF was considered to have a complex grammar for many interviewees. In other words, the structure of the stem was too complex and ambiguous for some test takers. This is pointed out below by a female test taker.

E5 "Item 2 had a very difficult structure to understand. I read the question several times but couldn't figure out what the question was about and how I was supposed to answer it"

INUEE reading items are all of multiple-choice format which is a potential source of bias for some test takers. Two female test takers expressed that the format of reading questions disadvantaged them because they believed that male test takers were better at answering such questions. This finding is in support of Aryadoust, Goh, and Kim's (2011) explanation for gender-ability DIF indicating that lower ability male test takers are more likely to attempt lucky guesses on multiple choice items. Similarly, Taylor and Lee (2012) stated that multiple choice reading comprehension items generally favored males while constructed-response items generally favored females. Also, Geramipour and Shahmirzadi (2019) found that on constructed response items female participants outperformed their male counterparts although the gap was not that large.

E6 "I am not good at answering multiple-choice reading questions. I can't understand why we always have to take such tests while there are so many ways to test our reading ability"

Passage length was another factor that male participants believed worked to their disadvantage. Two interviewees stressed that because the passages were too long, they ran out of time and could not perform well.

E7 "I couldn't finish reading passage 2 because it was too long and I ran out of time. I guess females have a higher chance of finishing the readings because they are generally better than males in speaking and reading quickly."

Finally, the reading question types were considered to be another reason why female test takers were not performing well on reading comprehension questions. Inference type questions were considered difficult for two female test takers who had experienced difficulty in comprehending implied meanings and making inferences. This finding aligns with Pae's (2004) study on gender differences in the reading comprehension subtest of Korean national entrance test for universities who concluded that item content which requires making a logical inference is against females. Moreover, he maintained that reading items that cover impression, mood, and tone are easier for males. Furthermore, Pae's (2012) seminal study over nine years indicated that item type is even a more reliable predictor of gender DIF than item content (i.e. passage topics) for the Korean EFL sample in his study.

E8 "item 11 was very hard for me because the answer was not directly mentioned in the passage. I think it favors males because they are better at understanding hidden meanings"

5. Conclusion and Implications

The first purpose of this quantitative and qualitative study was to investigate reading comprehension section of SET of INUEE for instances of DIF items through MH procedure. Moreover, since there is a paucity of research in the literature on DIF on causes of DIF from learner's perspective, 15 test takers were interviewed to elicit their views towards fairness of reading items on INUEE. The findings indicated that reading comprehension section only contained six gender DIF items. This a significant finding since reading skill contained very few DIF items while the assumption might be the reading skill shows more DIF items because the reading passages may come from different subject areas and function differentially for learners with different backgrounds. Moreover, the effect sizes of these six items were categorized as negligible which confirmed the fairness of reading section for both males and females.

Qualitative analysis of causes of DIF also confirmed the findings of quantitative part as test takers mostly considered INUEE a fair test. However, deep content analysis of test takers' attitudes towards underlying sources of DIF in the reading items revealed some common themes as potential sources of DIF. Interviewees expressed *topic familiarity, multiple-choice format of the test, topic interest, passage length, and complex structure of test items* as possible causes of DIF in INUEE.

The findings of this study could be informative especially to Iranian English teachers, considering the great impact or washback that SET of INUEE has on learners and teachers. The content and format of this test substantially influence the day to day practices of English learners and teachers to the extent that most learning and teaching activities are geared toward successful performance on this test instead of successful learning of the English language. This approach must be reversed and learning English for communicative purposes through effective methods must be the primary focus in high school classes.

Iranian high school students who are candidates for the foreign language majors are judged for their general English proficiency only on the basis of the results of the Special English Test. The presence of DIF within such a vital examination can liquidate the validity of the test since anything that weakens fairness harms the validity of a test (Xi, 2010). Therefore, it is recommended that testing instrument creators and policy makers of the National Organization for Educational Testing (NOET) of Iran conduct some precise and detailed DIF analysis of their high stakes tests every year and apply the findings of these studies in their testing procedure.

Most DIF studies have focused on item content in reading comprehension for possible sources of DIF while in this study it was shown that item type lead to differential performance as well. Based on his study on causes of gender DIF over nine years, Pae (2012) concluded that item type is a more reliable predictor of gender DIF than item content. Therefore, it is suggested that designers of INUEE and researchers address the impact of item type on test takers' performance in future studies.

References

Abdorahimzadeh, S. (2014). Gender differences and EFL reading comprehension: Revisiting topic interest and test performance. *System*, *42*(1), 70–80.

Alavi, S.M., Rezaee, A, & Amirian, S.M.R. (2011). Academic Discipline DIF in an English Language Proficiency Test. *Journal of English Language Teaching and Learning, 5* (7), 39-65

Amirian, S.M.R., Alavi, S.M., & Fidalgo, A.M. (2014). Detecting Gender DIF with an English Proficiency Test in EFL Context. *Iranian Journal of Language Testing, 4* (2), 187-203

Aryadoust, V. (2012). Differential Item Functioning in While-Listening Performance Tests: The Case of the International English Language Testing System (IELTS) Listening Module. *International Journal of Listening*, *26*(1), 40–60.

Aryadoust, V., Goh, C. C. M., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, *8*(4), 361–385.

Banerjee, H. L. (2016). Test fairness in second language assessment. *Working Papers in TESOL & Applied Linguistics*, 16(1), 54-59.

Barati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high-stakes tests: the effect of field of study. *IJAL, 19*(2), 27-42.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, SA: Sage.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Fidalgo, A, Alavi, S.M, & Amirian, S.M.R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing, 31(4),* 433-451

Flores, M.A., Simao A. M.V., Barros A., Pereira D. *(*2014*).* Perceptions of effectiveness, fairness and feedback of assessment methods: A study in higher education. *Studies in Higher Education, 40*(9), 1523-1534

Geramipour, M. (2020). Item-Focused Trees Approach in Differential Item Functioning (DIF) Analysis: A Case Study of an EFL Reading Comprehension Test. *Journal of Modern Research in English Language Studies, 7*(2), 123-147

Geramipour, M. Shahmirzadi, N. (2019). A Gender–Related Differential Item Functioning Study of an English Test. *The Journal of Asia TEFL*, *16*(2), 674-682

Hill, Y. Z., & Liu, O. L. (2012). Is There Any Interaction Between Background Knowledge and Language Proficiency That Affects Toefl Ibt ® Reading Performance? . *ETS Research Report Series*, *2012*(2), i–34.

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

Keshavarz, M. H., Atai, M. R., & Ahmadi, H. (2007). Content schemata, linguistic simplification, and EFL readers' comprehension and recall. *Reading in a Foreign Language*, *19*, 19-33.

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

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27* (2), 183-189.

Lam, T. C. (1995). *Fairness in performance assessment*. ERIC digest. Retrieved from: http://www.ericdigest.org/1996-4/fairness.html.

Lee, H., & Geisinger, K. F. (2014). The effect of propensity scores on DIF analysis: Inference on the potential cause of DIF. *International Journal of Testing, 14*, 313-338.

Li, L., X. Liu, & Steckelberg, A. L. (2009). "Analyzing Peer Feedback in a Technology-
facilitated Peer Assessment." In *Research Highlights in Technology and Teacher Education*, edited by C. D. Maddaux, 213– 221. Chesapeake, VA: AACE.

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). Does Content Knowledge Affect TOEFL IBT™ Reading Performance? A Confirmatory Approach to Differential Item Functioning. *ETS Research Report Series*, *2009*(2), 1-29.

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.

Pae, T. I. (2004). Gender effect on reading comprehension with Korean EFL learners. System, 32, 265–281.

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

Pae, T. & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, *23*(4), 475-96.

Park, G. P., & French, B. F. (2013). Gender differences in the Foreign Language Classroom Anxiety Scale. *System, 41*(2), 462–471.

Scott, S., Webber, C. F., Lupart, J. L., Aitken, N., & Scott, D. E. (2014). Fair and equitable assessment practices for all students. *Assessment in Education: Principles, Policy, & Practice, 21*(1), 52-70.

Suh, Y., & Talley, A. E. (2015). An empirical comparison of DDF detection methods for understanding the causes of DIF in multiple-choice items. *Applied Measurement in Education, 28*, 48-67.

Suskie, L. (2000) Fair assessment practices: Giving students equitable opportunities to demonstrate learning, *AAHE Bulletin*, *52*(9) 7-9.

Toker, D. (2019). Topic Familiarity Matters: A Critical Analysis of TOEFL iBT Reading Section. *TESL-EJ*, *23*(1), 1-9

Wu, A. D. & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*(3), 287-300.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147-170.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Summary* (p. 5). Ottawa, ON: Directorate of Human Resource Research and Evaluation, Department of National Defense.

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.

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.