AJELP: The Asian Journal of English Language & Pedagogy
ISSN 2289-8689 / e-ISSN 2289-8697, Vol 10 No.2 (2022),56-67

INVESTIGATING DIFFERENCES BETWEEN GENDERS IN AN ENGLISH ENTRANCE TEST IN A PUBLIC UNIVERSITY

Kamarul Ariffin Ahmad*

Faculty of Languages and Communication
Universiti Pendidikan Sultan Idris (UPSI), Malaysia

Izazol Idris

Faculty of Human Development
Universiti Pendidikan Sultan Idris (UPSI), Malaysia

email: kamarul.a@fbk.upsi.edu.my, izazol@fpm.upsi.edu.my

Published: 22 December 2022

To cite this article (APA): Ahmad, K. A., & Idris, I. (2022). Investigating Differences between Genders in an English Entrance Test in a Public University. *AJELP: Asian Journal of English Language and Pedagogy*, *10*(2), 56–67. https://doi.org/10.37134/ajelp.vol10.2.4.2022

To link to this article: https://doi.org/10.37134/ajelp.vol10.2.4.2022

Abstract: English entrance test is designed to select candidates to join English programmes that are offered in public universities. Many entrance test instruments have been developed to serve that purpose but there are quite a number of instruments which are developed by the lecturers are left without further investigation on how the instruments work to all the test-takers. This research analysed reading, vocabulary and grammar instrument that has been used in a public university in Malaysia as the entrance test for its English programme. The paper consists of 30 multiple-choice questions (10 questions for each section) and was administered to selected candidates who scored a minimum B+ grade in the *Sijil Pelajaran Malaysia* (SPM) examination. This research aimed to identify the reliability of the instrument, how each gender interacted with each of the items based on the comparison of the facility indices, to identify the DIF items and finally to classify them based on CEFR, Bloom's taxonomy and grammar topic. Based on the analysis with Winsteps, this research found that the reliability of the instrument is very good (*.98*), female test-takers are better when it comes to higher vocabularies in CEFR, male candidates perform better than female counterparts when it comes to preposition and tenses, female candidates perform well when it comes to higher rank of items based on Bloom's taxonomy, and there are three items flagged as showing DIF. This research concluded that, the instrument requires further improvements even though the reliability index is high. Some item which were flagged as showing DIF need to be reviewed and either be replaced or re-write.

Keywords: DIF, entrance test, gender, Rasch

INTRODUCTION

Fairness is a major issue when it comes to testing. An instrument which uses any items that exhibit unfairness properties may reduce its validity (Messicks, as cited in Sax, 2005; Seyed Mohammad Reza Amirian, Behzad Ghonsooly & Seyedah Khadijeh Amirian, 2020) thus making the interpretation and conclusion made of the score inaccurate (American Educational Research Association [AERA], 2014). Pertaining to the concerns, test developers have started to pay attention to test fairness as one of the properties to be looked at when they analyse test items. Generally, fairness is given an emphasis for tests that are utilised in making high-stake decision (Moghadam & Nasirzadeh, 2020).

Many experts and researchers suggest that those kinds of test undergo a proper procedure to offer the best possible fair treatment to the minority groups, who will be sitting for the test. Fairness is having a broad perspective and complex which not only concerns the content of the test but other aspects as well (Moghadam & Nasirzeh, 2020) such as the structure of the stem of the items, the testing procedure as well as the way one responds to the stimulus provided in the test. Even though, theoretically test developers or users should expect that the test takers respond differently, due to their different ability, fairness only happens when these differences are not consistent among a sub-group of test takers with the same ability (Boyer, 2020).

The subgroups can be defined based on the characteristics of the test takers which may include gender, ethnics and even education level such as undergraduates and postgraduates (Davies, 2010). These subgroups are often studied in test fairness research that involves the analyses on how an item works and the probability of getting the correct response which is also known as differential item functioning (Scott, et al., 2010; Chen & Revicki, 2014).

Differential Item Functioning (DIF)

Since the late 1960's, many methods on how to detect and eliminate items that exhibit DIF surfaced and according to Rezaee and Shabani (2010) it started when the test developers wanted to develop a fair test to white and African American test takers in the country. Since then, test developers have started to eliminate items that exhibited DIF in order to improve the test papers (Osterlind & Everson, 2009 in Shermis, Mao, Mulholland & Kieftenbeld, 2017) and DIF analysis has been done many fields including Mathematics, English and History (Siti Rahayah Ariffin, Rodiah Idris & Noriah Mohd Ishak, 2010). DIF refers to a test item that has a different interaction or different probability of getting the correct answers between two comparable subgroups such as between genders, sub-ethnic groups or even socio-economics status (AERA, 2014; Park, Pearson & Reckase, 2005; Shermis, et al, 2017). In the recent years, detecting items with differential item functioning has become the new customary practice in reducing biasness and is considered as one of the aspects to be studied in validating an instrument (Rezaee & Shabani, 2010) and an instrument that has been endorsed to be DIF free indicates high reliability (Siti Rahayah Ariffin, Rodiah Idris & Noriah Mohd Ishak, 2010).

Columbia University, who published the article Differential Item Functioning (2021), further explained that differential item functioning can be categorised into two which are benign and adverse. Benign differential item functioning is the representation of the characteristics of the subgroups that is not considered as measurement error. Whereas adverse differential item functioning is classified as measurement error and suggests that a particular subgroup experiences bias. On the other hand, differential item functioning could also be divided into two subcategories which are uniform DIF and non-uniform DIF.

Uniform differential item functioning refers to the probability of all the group members in a particular subgroup answering the item correctly or incorrectly as compared to the reference group. Their responses are systematic. Whereas in non-uniform differential item functioning, the response may be different among the group members in a particular subgroup and at some point, it matches the responses from the other group (read: reference group). This may happen when the high achievers and low achievers responded differently to an item. The debate about non-uniform DIF is ongoing and as the consequence, many ways are surfacing on how to assess and determine test fairness and the most usual and widely discussed subject matter in English language testing field would be differential item functioning (DIF) for gender (GDIF) (Rezaee & Shabani, 2010; Moghadam & Nasirzeh, 2020).

Siti Rahayah Ariffin, Rodiah Idris and Noriah Mohd Ishak (2010) further explained on how differential item functioning may happen among the test takers. Firstly, when an item is easy for a group of test takers and the same item is difficult to the other. Secondly, DIF happens when one group is performing as expected and the other is performing less than usual. Similarly, DIF may also happen when one group is performing as expected and the other is performing beyond what is expected. Lastly, DIF happens when an item is categorised as difficult item, but the same item is significantly more difficult for the other group. This is not to be confused with discrimination, as discrimination power provides insights about how well the item differentiate high achievers and low achievers, and DIF somehow discriminate between groups of the same characteristics.

Cauffman and Macintosh (2006) iterated that differential item functioning in Rasch model only measures uniform DIF and with the notion of the model, that computed the raw scores to logits to be mapped on a linear scale, Rasch views DIF as the location of an item on the scale differs between two groups. They further added that, items with DIF, if not taken seriously and it is consistent for the subgroup, will result in making problematic conclusion. This could be serious and misleading especially when it involves decision making as well as selection.

Differential item functioning not only being studied across subjects, but it is also being studied across groups of test takers. The use of DIF also expands to validation of health instruments or inventories which includes mental health (Cauffman & Macintosh, 2006). Some studies had made comparisons between ethnic groups as well as education background (Oliveri, Ercikan, Lyons-Thomas & Holtzman, 2016; Ercikan, Roth, Simon & Lyon-Thomas, 2014; & Oliveri & Zumbo, 2014) and they found that items could show DIF with the language of instruction of the schools. Apart from that, Oliveri et al (2016) have found that test-takers' socio-economic status (SES) may also contribute to DIF. Sandilands, Oliveri, Zumbo and Ercikan (2013) on the other hand found out that culture, and the translation and adaptation of the items may contribute to DIF.

Gender differences have been studied when it comes to measurement that went across many subjects including Mathematics, English and Sciences. Researchers have found that even high stakes tests, such as Programme for International Student Achievement (PISA) and Trends in International Mathematics and Science Studies (TIMSS), could contain GDIF items (Le, 2009; Kan & Bulut, 2014; Hauger & Sireci, 2008). Interestingly, Le (2009) found that many items for language section were flagged with GDIF based on his research for PISA which involved 83,000 students across 50 countries. With this kind of test, analyses with DIF are crucial so that the items, which would be added to the bank for the subsequent use, might not be dropped and with a careful review, they would be reliable and valid (Le, 2009).

The previous research has shown that DIF could happen across many sub-groups but due to certain circumstances, this study could only be performed with gender. It is also wise to note that the enrolment to the course consists of more female students as compared to the male counterpart. This raises the question whether the items in the entrance test could favour female candidates. Therefore, this study was conducted and aimed to answer the following questions:

i. Is there any difference of difficulty indices between male and female test-takers?
ii. Are there any items that shows DIF?
iii. If there is any, what is/are the possible reason(s) of the DIF?

BACKGROUND OF THE STUDY

English Entrance Test

English entrance test is common in many countries where English is not their spoken language. In general, English entrance test would measure candidates' linguistic abilities to go through and complete a course of a study which uses English as the medium of instruction. In this study, the English entrance test is used as selection thus the scores from the test will facilitate the faculty in decision making. The candidates who applied are carefully selected to sit for the test and only who scored B+ and above in English subject for *Sijil Pelajaran Malaysia* would be called for the test. The test is in nature a proficiency test which comprises of three components which are Reading, Vocabulary and Grammar.

Coombe, Folse and Hubley (2007) explained that language proficiency test is a test that collects data on the ability of the candidates at different level of use of the language. While Peng, Yan and Cheng (2020) defined proficiency as the ability for someone to use language skills such as reading, listening and writing in communication, Coniam and Palvey (2013) added two more dimensions to be added in the construct of proficiency which are grammar and vocabulary. In this English Entrance Test, all skills suggested by Peng, Yan and Cheng (2020) and Coniam and Palvey (2013) are taken into account. Despite the skills listed, only an instrument in the test battery that consists of Reading, Vocabulary and Grammar involved in this study.

Previously, the idea of the inclusion of grammar and vocabulary was debatable and was deemed as inappropriate for a proficiency test (Kamarul Ariffin Ahmad, Muhamad Lothfi Zamri and Nora Liza Abdul Kadir (2015) but Coombe and Davidson (2014), and Zhao and Liu (2019) opposed to this idea by stating that grammar and vocabulary are needed vital knowledge for one to function in delivering messages and indeed crucial to be tested in proficiency tests. Each of the sections in the instrument consists of ten multiple choice items. For Reading, a non-linear text is presented to the students and is accompanied with ten multiple choice questions. For Vocabulary and Grammar sections, a hybrid of a supply response format and multiple choice items is used where there will be ten blanks in a reading passage; each for each section.

Testing of Grammar

Grammar knowledge could be considered as one's technical knowledge on how words are arranged and used to deliver a message. As has been mentioned earlier, testing grammar is debatable as it is deemed to be testing lower level of knowledge which does not require much of thinking but rather memorising (Gezer, Oner Sunker & Fahin, 2014). Given that the instrument is used for selection, the inclusion of grammar is seen as needed to ensure those

who are selected are the cream of the crop. Kamarul Ariffin Ahmad, Muhamad Lothfi Zamri and Nora Liza Abdul Kadir (2015) further added that developing multiple choice items to test grammar could be challenging as grammar items, due to its technicality, tend to provide giveaway to the test takers. Therefore, the development of the items has to be very careful and meticulous to ensure that giveaway can be avoided thus sustaining the reliability and validity of the items.

Testing Reading and Vocabulary

Testing of reading was considered as the easiest part of testing the language by many researchers in those days (Kitao & Kitao, 1996) and according to Alderson (2000), there is no specific ways a test developer can adopt to test reading; it could be verbally, written and with just a choice of options. They only put an emphasis on the selection of text when it comes to reading test, where the selection would be heavily depended on the objective of learning the language. For instance, if the purpose of learning the language is for academic purposes, the chosen text has to be related or at least a representation of an academic writing.

When it comes to the development of the items, many developers use Bloom's taxonomy for cognitive test. What is important during the development of the items for a reading test is that the items have to be the representation of the test takers' reading ability (Arung, 2013). In addition to that, Arung also suggested that the item developer to consider the length of the text with a notion that, those who manage to answer complete the reading test are ought to obtain the desired outcome of the instruction. Salvia and Ysseldyke (2001) summed up the whole idea of a reading test with suggestions that a reading test instrument should begin with the simplest of items and tasks for reading and the difficulty of the items is increasing as the test progresses to the end.

Testing vocabulary could be done either in isolation or in context (Nation & Chung, 2009) and Jones (2004) claimed that many vocabulary tests that were conducted in isolation used recognition and recall method with several formats such as using pictures and annotations. Besides being tested in isolation, Nation and Chung (2009) stated that vocabulary testing is common in proficiency, diagnostic and achievement tests. They added that tests for vocabulary usually inclusive of either breadth of vocabulary which is similar to vocabulary size or depth of vocabulary which involves testing the knowledge and accuracy of words used in a language.

On the other hand, Webb and Sasao (2013) iterated that the development of vocabulary tests has a small progress for about 30 years of their research and suggested that this is due to the little contribution of researches towards vocabulary assessment. This results in many vocabulary tests to use the approach and format that is deemed to be appropriate for the instructional objectives. Alderson, Clapham and Wall (2010) and Salvia and Ysseldyke (2001) explained that the most favourable approach and format for this purpose is gap-filling; similar to what is adopted in the instrument studied in this research.

Sources of Differential Item Functioning

A study to identify, review and retest DIF items may be longitudinal and rare but it is valuable (Park, Pearson & Reckase, 2005). Accordingly, reasons for DIF may differ from a subject to another or even the format of the test itself; this happens between both genders, language spoken at home (mother tongue), countries as well as culture. Shermis et al. (2017) listed three possible sources of DIF and they are 1) the clarity of the stem of the items and sensitivity toward unique culture of the test-takers, 2) the linguistic nature of the items, and 3) the organization of the items in the test.

At times, the verbosity of the stem may hinder the intended purpose thus confusing the test-takers. A stem of an item must be concise and clear which means it should not require the test-takers to spend time thinking of how to interact but interacting with the item with the knowledge they possess. Certain tests which may be developed to be used locally may contain colloquial words which are not universally known, must be carefully reviewed before it is used with students from a different culture. When it comes to the format of the items, a study by Le (2009) found that multiple choice questions seem to favour males whereby females perform better in open-ended questions.

In terms of subjects, Becker (1989) Jovanovic, Solano-Fores and Shavelson (1994), Young and Fraser (1994) and Burkam, Lee and Smerdon (1997) as cited in Le (2009) said that male test-takers perform better when it comes to physical, earth and space science topic, and this is consistent with Rezaee and Shabani (2010) who claimed that male test takers perform better in technical and sciences subjects. Khorramdel et al (2020) on the other hand found that male students perform better in Mathematics while female students do well in reading especially when it comes to cognitively demanding tasks. They also found that the gap between the genders was significantly higher in reading than in Mathematics. Interestingly, in Hasni Shamsuddin et al (2020) study, they found that when it comes to Mathematics test, female students perform better at straightforward items where male students tend to answer correctly when the items are more complex and require more understanding of the problem presented.

Both genders generally will have differences when it comes to language test and this should not be neglected because according to Seyed Mohammad Reza Amirian, Behzad Ghonsooly & Seyedah Khadijeh Amirian "gender DIF is a major source of bias that can threaten the validity of a standardized test" (p.90, 2020). In addition to that, Scheuneman and Grima (1997; as cited in Kan & Bulut, 2014) found that linguistic properties of an item may influence the way the items function for female and male test-takers. Linguistics research have proven since long ago that both genders have their own unique and specific ways of receiving and delivering the language (Shermis, et al., 2017) yet the DIF studies involving reading comprehension items are rare (Taylor & Lee, 2012). In reading comprehension test, male and female test-takers are said to have different interactions toward concrete and abstract subject (Wedman, 2018) where female test-takers are having advantages toward items or stimuli containing human subject, emotions and with aesthetic values.

Methods on Detecting Differential Item Functioning

There are many ways that a researcher can employ to detect items with DIF and these methods may differ for dichotomous and polytomous items. Ultimately, all methods that were developed by the researchers, assume that the test-takers, regardless of their background, have the same opportunity to answer the items correctly. According to Le (2009), the popular methods include Mantel-Haensazel by Dorans and Holland (1993), logistic regression by Swaminathan and Rogers (1990), standardisation by Dorans and Holland (1993) and Dorrans and Kulick (1986), confirmatory factor analysis and multidimensional approach by Joreskog and Sorbom (1989), and also item response theory (IRT) with the use of Rasch model (Sumintono & Widhiarso, 2015). Siti Rahayah Ariffin, Rodiah Idris and Noriah Mohd Ishak (2010) supported the notions and claimed that in general the three ways of detecting item with differential item functioning as mentioned as the most popular methods. On the other hand, Hambleton, Swaminathan and Rogers (1991; as cited in Erciken, et al., 2014) stated that an item can be considered as having DIF even when it has different difficulty and discrimination indices between two sub-groups.

In Rasch model, there are two assumptions that are needed to be fulfilled in order for further investigation, in this case the DIF, is to be conducted and they are the unidimensionality and the local independence. After the assumptions are fulfilled, further investigation with Rasch model can be conducted. Consequently, there are two ways of viewing and investigating instrument for DIF can be adopted in Rasch model which are prior to the test, this takes place during the item development phase, and the post test, this involves the statistical procedures in detecting DIF items. Sumintono and Widhiarso (2015) had listed a list of questions to help item developers in detecting and reducing DIF prior to the tests:

i. does this test have different format in its batteries to ensure the test-takers with different learning styles have equal opportunities?
ii. has the developer taken into considerations about the test-takers' gender, social economic status, ethnic or culture based on previous research?
iii. do all test-takers aware on how they should respond to each of the item (format)?
iv. has the test developer considered the previous results from previous tests?
v. do the test-takers know why they learn and how are they being assessed?
vi. does the test developer clear with the relationship between learning and assessment? and
vii. is the test developer sure that this test will not insult nor judge a sub-group?

Given that the test instrument was already developed and administered, this research will proceed to the post test procedures where the responses will be collected and empirically analysed. Cauffman and Macintosh (2006) explained the formula of computing DIF which is used in the Winsteps software:

$$t = \frac{d_1 - d_2}{\sqrt{var\ d_1 - \ var\ d_2}}$$

In this formula, d_1 and d_2 are the locations of the two groups on the scale.

By using the software for Rasch model such as Winstep, according to Eakman (2012; cited in Hasni Shamsuddin et al, 2020) unidimensionality is achieved when the eigenvalue of the principal component analysis (PCA) is less than *3.0* and the percentage of the unexplained variance must be below *10%*. For local independence, Aryadoust (2018) suggested that, with the analyses computed from Winstep, any item that correlates with each other with *>.60*, can be classified as not fulfilling the assumption of local independence.

In order to identify items with DIF, the researcher have to look for items with $p<0.5$ (Sumintono & Widhiarso, 2015) and the size of the DIF can be categorised as moderate for ≥ 0.43 logits and large DIF size if the logits are ≥ 0.64, whereas logits below *0.43* can be considered as trivial of negligible for DIF (Zwick et al, as cited in Linacre, 2021). Similarly, Bond and Fox (2015) suggested that $p<0.5$ plus the *t* value of $\pm$ *2.00*.

METHOD

Sample and Instrument

The sample of this study consisted of 303 test-takers who had been selected nationwide to sit for the entrance test for an English programme at diploma level. All test-takers were those who applied to be enrolled and selected based on their *Sijil Pelajaran Malaysia* (SPM) English paper results where they achieved a minimum grade A-. The sample consisted of 171 female

and 132 male test-takers came from various states in Malaysia. The samples were called to take the test at various centres located in peninsular Malaysia and East Malaysia.

The test-takers were given 1 hour to complete 30 multiple-choice questions and one essay item. The only items taken into consideration for this study were the multiple-choice questions. These 30 items were divided into three categories which were vocabulary items for first category, grammar items and reading comprehension questions; for each category, there were 10 questions. Due to the secrecy act, the questions for the test are not to be disclosed.

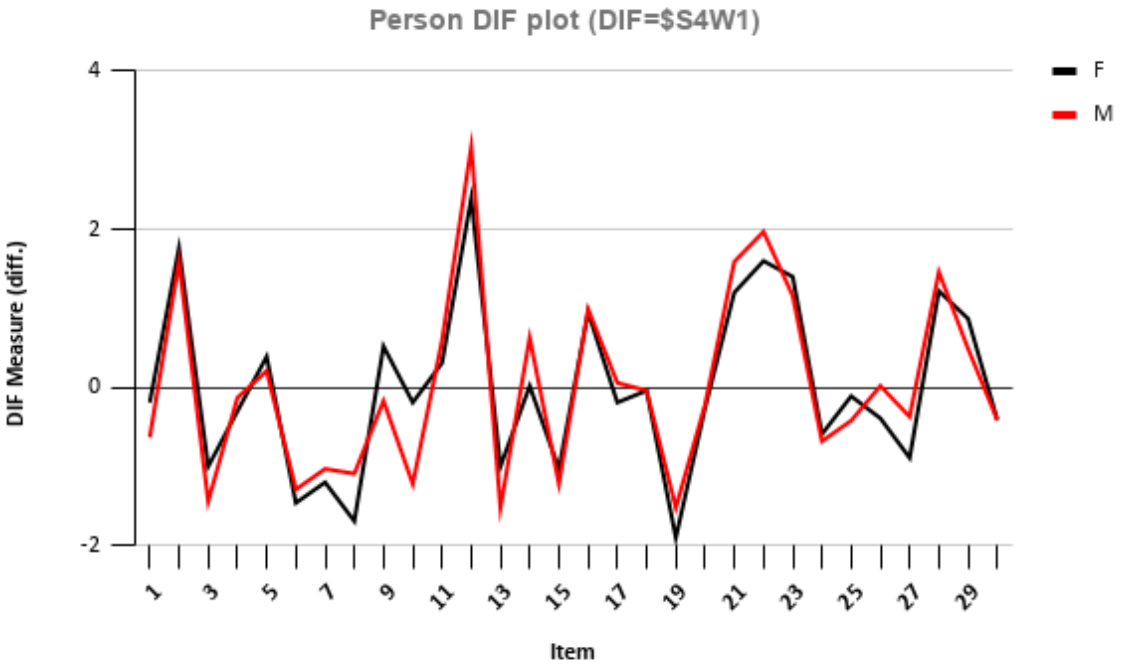
Data Analysis

The items in this study will be analysed with Rasch model by using Winsteps software and prior to performing the analysis to detect items with DIF, the reliability of the items will be obtained. The item fit will then be checked because if the items are fit in the model, the assumption of unidimensionality can be achieved. The analyses will then was continued by calibrating the items where the difficulty and discrimination indices for both genders were obtained, and it proceeded to obtaining and detecting the DIF items.

FINDINGS AND DISCUSSION

To begin with, the items had gone through the unidimensionality analysis where the items fulfilled the unidimensional assumptions with *5%* of the unexplained variance (must be *<10%*) and *2.0* for PCA residuals (must be *<3.0*). The analyses were continued with an item analysis to determine the difficulty indices for the items, the DIF size and the *p*-value for each of the items. The *p*-value is set at $p<.05$ to determine and item is having a statistically significant DIF.

Graph 4.1 *Graphic Comparison of Item Difficulty Indices between Male and Female Test-takers in an English Entrance Test for Vocabulary, Grammar and Reading Comprehension.*



Looking at the line graph and to answer the first research question, we can see that the logits for difficulty for both genders are showing differences across many items in the instrument. Nevertheless, this study also found that all items went on the same direction for both genders which means that both genders may have similar ideas when it comes to viewing an item as difficult or easy. At this level of analysis, one item indicated that it needs to be reviewed and that is item number 12. For both genders, item number 12 is beyond the test-takers capabilities and is outside of acceptable difficulty logits as suggested by Bond and Fox (2014) which is between *-2.00* to *+2.00*. Considering that this is an entrance test, the item developer may consider retaining this item to indicate that the selected candidates possess an extra knowledge as compared to rest. Nevertheless, a review is still suggested to eliminate other possibilities such as the verbosity of the stem, malfunctioned distractors or the content could be out of context.

The DIF analysis (in Table 4.1) showed that three items are flagged as statistically significant with DIF and they are item 9, item 10 and item 14. According to Zwick, Thayer and Lewis (1999) in Rasch analysis for DIF, a logit size of *<.43* can be considered as trivial and can be ignored (Aryadoust, 2017), logits between *.43* to *.63* can be considered as moderate and logits *.64* and above can be considered as large. Based on this notion, all three items needed reviews since all the three items fell into large effect DIF category.

Based on the difficulty logits for these three items, this study concluded that even though the direction of the graph went to a same way, item 9 showed that the item is considered hard for female test-takers (*.52*) and easy for the male counterparts (*-.17*) and therefore, this study concluded that this item should be dropped and replaced. Unlike the other two items, item 10 and item 14, both items are either hard or easy for both genders. Therefore, this study suggested that these two items needed to be reviewed in terms of the structure of the stem, distractors as well as the content.

Table 4.1 *Item analysis: Difficulty Index between genders, DIF Size and p-value for an English Entrance Test for Vocabulary, Grammar and Reading Comprehension.*

Item Number	Difficulty index		DIF Size	*p* value
	Female	Male		*<.05*
1	-.19	-.63	.51	.07
2	1.76	1.59	.38	.41
3	-1.00	-1.44	.15	.09
4	-.30	-.13	.23	.53
5	.39	.21	.15	.52
6	.39	.21	.12	.66
7	-1.46	-1.29	.71	.42
8	-1.69	-1.09	1.28	.07
9	***.52***	***-.17***	***.56***	***.01***
10	***-.19***	***-1.22***	***.58***	***.00***
11	.31	.59	.08	.28
12	2.39	3.04	.27	.10
13	-1.00	-1.52	.43	.08
14	***.02***	***.63***	***.70***	***.01***
15	-1.03	-1.22	.22	.57
16	.95	.99	.00	.82
17	-.19	.06	.37	.22
18	-.04	-.04	.04	.97
19	-1.90	-1.52	.31	.45

20	-.26	-.22	.21	.76
21	1.20	1.59	.31	.22
22	1.60	1.97	.06	.21
23	1.40	1.16	.28	.34
24	-.59	-.68	.26	.53
25	-.11	-.42	.27	.42
26	-.39	.02	.51	.12
27	-.89	-.37	.54	.09
28	1.22	1.46	.03	.39
29	.87	.49	.52	.12
30	-.41	-.41	.15	.70

A review of item number 10, which is vocabulary item, showed that male candidates were having the advantage over this item. This item has four distractors ranging from level B2 to C2 in CEFR. Meanwhile, for item number 14 which is Grammar item, female test-takers had the advantage to getting the correct answer and this item was about tenses (simple present tense). This maybe consistent with findings from Barati and Ahmadi (2010; as cited in Seyed Mohammad Reza Amirian, Behzad Ghonsooly & Seyedah Khadijeh Amirian, 2020) where they claimed that male test-takers perform better in vocabulary items and female test-takers are having advantage when it comes to Grammar. Nevertheless, this study is still inconclusive because these two items cannot be used to represent whether both genders possess a certain characteristic when it comes CEFR vocabulary levels as well as Grammar.

Judging from the verbosity of both items, item 10 used simple sentence and direct stem whereby item 14 had a complex structure of sentence in its stem. The context of both items also provided some insights where for item number 10, the topic was about heart disease (concrete and scientific topic), and for item number 14, the topic is about teaching and learning (aesthetic values). These are consistent with the findings from Wedman (2018), where he said that female test-takers have the advantage when it comes to complex sentence structures as well as topic aesthetic values, and male test-takers are having the advantage when it comes to simple, direct sentence structure with concrete or scientific topic. Similarly in Seyed Mohammad Reza Amirian, Behzad Ghonsooly & Seyedah Khadijeh Amirian (2020) study, they found that male tend to do better when it comes to science-related subject (in this case heart attack) whereas female tend to do better in humanities-related subject (in this case teaching and learning).

CONCLUSION AND RECOMMENDATION

The study of DIF has been a debate among test developers and is given attention when test biasness is discussed. Despite having few substantial sub-groups which may advantage than the other, some tests were developed without proper analyses and documentation for its properties, which includes DIF. Although this study may support previous research on one part, the other part which may provide some insights needs to be studied further – vocabulary (CEFR based) and Grammar. This could be due to the small number of items for each category which is used in this study – the entrance test uses one instrument for few years. Therefore, for future researches, this study suggest a bigger number of items so that the data could be conclusive when it comes to vocabulary and Grammar items.

REFERENCES

Alderson, J. C. (2000). Techniques for testing reading. *Assessing Reading*, 202-270. doi:10.1017/cbo9780511732935.008

Alderson, J. C., Clapham, C., & Wall, D. (2010). *Language test construction and evaluation*. Cambridge: Cambridge Univ. Press.

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

Amirian, Seyed Mohammad & Ghonsooly, Behzad & Amirian, Seyedeh. (2020). Investigating Fairness of Reading Comprehension Section of INUEE: Learner's Attitudes towards DIF Sources. *International Journal of Language Testing, 10*(2), 88-100.

Arung, F. (2013). Testing Reading. 10.13140/RG.2.1.4171.8484.

Aryadoust, V., 2017. *Rasch Measurement using WINSTEPS*. [video] Available at: https://www.youtube.com/watch?v=FDUYm7ZhXkw&t=533s

Aryadoust, V. (2018, Jan 21). *Rasch Measurement Unidimenstionality and Local Independence (Part 2)* [video]. Youtube. https://www.youtube.com/watch?v=2PgOxMy54iQ

Bond, T., & Fox, C. (2015). *Applying the Rasch Model Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Boyer, M. (2020, October 1). Accountability. Retrieved from https://www.nciea.org/blog/educational-assessment/fairness-educational-testing

Cauffman, E., & Macintosh, R. (2006). A Rasch Differential Item Functioning Analysis of the Massachusetts Youth Screening Instrument. *Educational and Psychological Measurement, 66*(3), 502-521. doi:10.1177/0013164405282460

Chen WH., Revicki D. (2014) Differential Item Functioning (DIF). In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_728

Coniam, D., & Falvey, P. (2013). Ten years on: The Hong Kong Language Proficiency Assessment for Teachers of English (LPATE). *Language Testing, 30*(1), 147-155. https://doi.org/10.1177/0265532212459485

Coombe, C., Folse, K. & Hubley, N. (2007). A Practical Guide to Assessing English Language Learners. Ann Arbor: The University of Michigan Press

Coombe, C., & Davidson, P. (2014). Common Educational Proficiency Assessment (CEPA) in English. *Language Testing, 31*(2), 269-276. https://doi.org/10.1177/0265532213514530

Davies, A. (2010). Test fairness: A response. *Language Testing, 27*(2), 171-176. doi:10.1177/0265532209349466

Differential Item Functioning. (2021, September 30). Retrieved from https://www.publichealth.columbia.edu/research/population-health-methods/differential-item-functioning

Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education, 27*(4), 273–285.

Gezer, M., Oner Sunkur, M. & Sahin, F. (2014). An Evaluation of the Exam Questions of Social Studies Course According to Revised Bloom's Taxonomy. GESJ: Education Science and Psychology, 28(2), 3–17.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing, 8*(3), 237–250.

Jones, L. (2004). Testing L2 vocabulary recognition and recall using pictorial and written test items. *Language Learning & Technology, 8*(3), 122-143.

Kamarul Ariffin Ahmad, Muhamad Lothfi Zamri & Nora Liza Abdul Kadir. (2015). An Investigation of the Frequency of HOT and LOT of Bloom Taxonomy in the Diploma English Entrance Exam. *AJELP: Asian Journal of English Language and Pedagogy, 3,* 228-241.

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.

Khorramdel, Lale & Pokropek, Artur & Joo, Seang-Hwane & Kirsch, Irwin & Halderman, Laura. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. Psychological Test and Assessment Modeling. 62. 179-231.

Kitao, S. K., & Kitao, K. (1996). *Testing Reading* (ED398258). ERIC.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, *9*(2), 122–133.

Linacre, M. (2021). (n.d.). *Table 30.1 Differential item functioning DIF pairwise*. Winsteps.Com.

Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia*, *10*(1). doi:10.1186/s40468-020-00105-2

Nation, P., & Doughty, C. (2009). Teaching and Testing Vocabulary. In T. Chung & M. Long (Eds.), *The Handbook of Language Teaching* (pp. 543-559). doi:10.1002/9781444315783.ch28

Oliveri, María Elena, Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286–300.

Oliveri, Maria Elena, Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education*, *29*(1), 17–29.

Park, H.-S., Pearson, P. D., & Reckase, M. D. (2005). Assessing the effect of cohort, gender, and race on differential item functioning (dif) in an adaptive test designed for multi-age groups. *Reading Psychology*, *26*(1), 81–101.

Peng, Y., Yan, W., & Cheng, L. (2020). Hanyu Shuiping Kaoshi (HSK): A multi-level, multipurpose proficiency test. *Language Testing*, *38*(2), 326-337. https://doi.org/10.1177/0265532220957298

Rezaee, A., Shabani, E. (2010). Gender Differential Item Functioning Analysis of the University of Tehran English Proficiency Test. *Research in Contemporary World Literature*, 14(56).

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment*. Boston, MA: Houghton Mifflin.

Shamsuddin, Hasni, Abd. Razak, Nordin, Thien, Lei Mee, Khairani, Ahmad Zamri. (2020). Do boys and girls interpret mathematics test items similarly? Insights from Rasch model analysis. *Asia Pacific Journal of Educators and Education*, *35*(1), 17–36.

Shermis, M. D., Mao, L., Mulholland, M., & Kieftenbeld, V. (2017). Use of automated scoring features to generate hypotheses regarding language-based DIF. *International Journal of Testing*, *17*(4), 351–371.

Siti Rahayah Ariffin, Rodiah Idris, & Noriah Mohd Ishak. (2010). Differential Item Functioning in Malaysian Generic Skills Instrument (MyGSI). *Jurnal Pendidikan Malaysia*, *35*(1), 1-10.

Sumintono, Bambang & Widhiarso, Wahyu. (2015). Aplikasi Pemodelan Rasch pada Assessment Pendidikan.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246–280.

Webb, S. A.; Sasao, Y. (2013). New Directions In Vocabulary Testing. RELC Journal, 44(3), 263–277. doi:10.1177/0033688213500582

Wedman, J. (2018). Reasons for gender-related differential item functioning in a college admissions test. *Scandinavian Journal of Educational Research*, *62*(6), 959–970.

Zhao, C., & Liu, C. (2019). An evidence-based review of Celpe-Bras: The exam for certification of proficiency in Portuguese as a foreign language. *Language Testing*, *36*(4), 617-627. https://doi.org/10.1177/0265532219849000