International Journal of Learning, Teaching and Educational Research
Vol. 19, No. 6, pp. 27-45, June 2020
https://doi.org/10.26803/ijlter.19.6.2

Examining the Fairness of Language Test Across Gender with IRT-based Differential Item and Test Functioning Methods

Burhanettin Ozdemir
Prince Sultan University, Riyadh, Saudi Arabia
https://orcid.org/0000-0001-7716-2700

Abdulrahman Hadi Alshamrani
National Center for Assessment, Riyadh, Saudi Arabia
https://orcid.org/0000-0002-6560-3422

Abstract. Test fairness is an important indicator of the validity of test results. The fairness and equity require ensuring that the background characteristics of test-takers, such as ethnicity and gender, do not affect their test scores. Differential item functioning (DIF) methods are commonly used to detect potentially biased items that lead to the unfair assessment of the performance of test-takers with the same ability levels coming from the different cultural, social, demographic, and linguistic backgrounds. This study aims at detecting potentially biased items across gender and examining their effect on test scores to ensure the fairness of test results for each domain and the entire test. Item response theory (IRT) based Lord's chi-square DIF method at item level and Mantel-Haenszel/Liu-Agresti differential test functioning (DTF) method at test level were implemented to the English Placement Tests (EPT) administered to high school graduates by the National Center for Assessment. The results show that 6 items of the EPT exhibit DIF for the entire test. Two of them are related to reading comprehension and four to the structure domain, while none of the compositional analysis methods shows DIF. These results indicate the existence of content specific DIF effect. Additionally, two items exhibit uniform DIF, one of which shows DIF favoring male students and the offer favoring female students. The small to moderate DTF effect associated with sub-domains and the entire test imply that DIF effects cancel each out, assuring the fairness of results at test level. However, the items with substantially high DIF values need to be examined by content experts to determine the possible cause of DIF effects to avoid gender bias and unfair test outcomes. We also suggest conducting further studies to investigate the reasons behind the content specific DIF effects in language tests.

Keywords: test fairness and validity; gender bias; language testing; differential item functioning; differential test functioning

1. Introduction

The major concern of the stakeholders in education and test-takers is to ensure the fairness tests. The best way to provide fairness regarding the decision made upon a test is to increase the validity and reliability of test results. Therefore, any effort to minimize confounding factors such as random and systematic errors, and increase validity and reliability of test will serve the purpose of developing fair tests and valid test scores for examinees belonging to different groups. Examining the factorial structure of a test and differential functioning at the item level and test level are commonly used methods to assess the reliability and validity of test scores. Differential functioning may occur when items and tests produce different results for different groups consistently and therefore lead to invalid test scores and decisions made based on these scores.

The stakeholders that take part in educational test development and assessment processes explicitly emphasize the importance of fairness in test results regarding different subgroups. They put a substantial amount of effort to detect irrelevant factors threatening the construct validity of the test. They are aware of the necessity and importance of collecting evidence to justify the validity and fairness of the tests and change the testing policies accordingly. Recently, the European Federation of Psychological Association has proposed a model for collecting evidence of construct validity (Evers, Muñiz, Hagemeister, Høstmælingen, Lindley, Sjöberg, & Bartram, 2013; Hope, Adamson, McManus, Chris, & Elder, 2018) in which using differential item functioning (DIF) is considered as an important method for assessing the quality of the test. Moreover, the Test Commission of the Spanish Psychological Association has emphasized the critical role of DIF analysis in the context of test fairness (Hernández, Tomás, Ferreres, & Lloret, 2015; Hope et al., 2018).

A common definition of differential item functioning (DIF) is that an item is said to exhibit DIF when the probability of correct response to an item differs across subgroups with the same ability level (Hambleton & Rogers, 1989). DIF types are classified into two groups that are uniform DIF and non-uniform DIF. An item exhibits uniform-DIF when the difference in item performance is consistent and in favor of certain subgroups across the entire range of ability. However, if this difference between subgroups is not consistent, then DIF is identified as non-uniform (Hambleton, Clauser, Mazor & Jones, 1993).

The existence of DIF is an indicator of item bias and the presence of the secondary latent trait besides the primary latent trait that an item aims to measure. However, this secondary latent trait does not always imply bias or cause unfair assessment. If the secondary latent trait is related to the primary trait and occurs due to the nature of the measured structure, then the item is not labeled as unfair regardless of the differing performance of sub-groups. This situation was illustrated in a study conducted by Drabinová and Martinková (2016). They found that one DIF item related to childhood illness in which females showed better performance than males. However, a detailed investigation of content experts revealed that this performance difference occurred since women are more experienced than men since they spend more time with their children in the Czech Republic (Martinková et al., 2017).

Therefore, the performance difference between women and men in this example reflects the true ability difference and does not cause unfairness. Therefore, an item may display DIF, however, this finding does not provide enough evidence to classify this item as a biased item. Bias is related to systematic error in test administration and contents and relies on both statistical tests and expert opinions (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Wiberg, 2006), while DIF only relies on statistical tests.

Although there are many different parametric and non-parametric methods to detect DIF, which method to utilize is a main concern of researchers, since each method has advantages and shortcomings (Anastasi & Urbina, 1997; Hunter, 2014). For instance, some methods fail to detect non-uniform DIF but effective when the sample size is small, such as the Mantel Haenszel and Rasch methods, while some methods are capable of detecting non-uniform DIF but requires large sample size (Ferne & Rupp, 2007; Lai, Teresi, & Gershon,2005) such as IRT based Raju's area method and Lord's chi-square method. These aforementioned methods are the most commonly used exploratory methods utilized to identify differential item functioning for categorical variables that represent existing subgroups such as gender, nationality, and age groups (Aryadoust & Zhang, 2016). The next step after detecting items exhibiting DIF is to investigate the possible source and cause of occurrence of DIF (Zhu & Aryadost, 2020).

The EPT is administered to high school graduates by the National Center for Assessment (NCA) in Saudi Arabia. The results of the EPT has been used by several colleges, universities, and institutes to measure students' language skills, to screen their improvements across different levels or to determine their required language proficiencies (Education & Training Evaluation Commission [ETEC], 2020). Luo and Al-Harbi (2016) examined the factorial structure of the EPT with unidimensional and DIMTEST methods. They found strong evidence supporting the unidimensionality of the EPT which justified the usage of the IRT-based models instead of the classical test theory method (CTT).

1.1. Literature review

There have been many studies conducted to ensure test fairness across different subgroups and to define the potential source of DIF effects. Drabinová and Martinková (2016) conducted a study to determine the potential sources of DIF effects concerning the presence of the secondary latent trait besides the primary latent trait. They concluded that the existence of the DIF effect for some items did not mean that these items were biased because some of the DIF effects reflected the relationship between the secondary latent trait and the primary latent trait. Thus, they suggested that one should avoid labeling DIF items as biased items without the investigation of the content experts.

Chubbuck et al. (2016) studied DIF effects in the context of differing contents across gender groups. They employed the Mantel–Haenszel and standardized DIF methods to detect DIF items for each content domain. They found that the males showed better performance than females in reading comprehension items. They also defined the lack of sufficient context in the sentence completion items as a potential source of DIF effects. Finally, they recommended utilizing more than one DIF methods to increase the accuracy of the results. Wedman (2018)

examined if the language ability of non-native test takers that took the test in a language other than their mother tongue affected their performance compared to the native speakers. It was found that the deficiency in the language skill of non-native test takers caused the DIF. Moreover, He defined the failure in wording the content clearly in an item as a potential source of DIF effects. (Siegel, 2007, Wedman, 2018).

In one study, Stage (2005) investigated the SweSAT test items administered in spring concerning DIF across gender groups. The Mantel–Haenszel DIF method was employed to detect DIF items and It was found that 21 out of 122 items exhibited DIF across gender groups. Among these DIF items, 10 items related to the quantitative and verbal domains were in favor of female students. However, this study did not find any patterns among DIF items and did not suggest anything about the potential source of DIF effects. Federer and her colleagues (2016) employed the Mantel-Haenszel DIF method for detecting potential DIF items in the context of natural selection across gender groups. They specifically focused on open-ended questions. It was found that women outperformed men for the items that require applying the knowledge to the new conditions. Admitting the fact that the developed measurement instrument showed gender bias and, they did not suggest anything about the potential source of DIF effects due to the complex nature of DIF structure.

Similarly, Lin and Wu (2003) used DIF and differential bundle functioning (DBF) to detect items that function differently across gender on the EPT administered to Chinese EFL learners. For this purpose, they used the SIBTEST methods to detect DIF items. The results of this study indicated that the testlets (item bundles) containing the listening comprehension items showed DIF in favor of females, while the testlets containing the grammar and vocabulary exhibited DIF in favor of males. Thus, these findings provide strong evidence about content specific DIF. Pae (2012) studied the trends in the magnitude of DIF on the English subtest administered to the Korean students across gender groups for the nine-year period. He used the Mantel-Haenszel and IRT-based likelihood ratio test methods to detect the DIF items. Moreover, the study examined the effects of reading strategies and perceived interest on the magnitude of DIF. The results of this study showed the strong evidence about the relationship between the type of items and DIF, and a substantial interaction between the test takers interest in the items and the magnitude of DIF across gender.

It is substantially important to run DTF analyses along with DIF since items are small and unreliable compared to the test (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001) and the total amount of DIF provides an overall effect of DIF on test scores even when there is no item detected as DIF in a test (Hunter, 2014; Shealy & Stout, 1993). Additionally, DTF values can be negligibly small when these DIF items are in favor of different subgroups or in a different direction where DIF effects cancel each out (Borsboom, 2006; Zhu & Aryadoust, 2020). DTF is also important since decisions about examinees are not made at item-level, but test-level (Ellis & Raju, 2003; Roznowski & Reith, 1999; Pae & Park, 2006; Zumbo, 2003). More detailed information about the DIF and DTF methods is provided in the following sections.

1.2. IRT based DIF methods

The IRT based DIF methods are suggested in the case of a large sample size. The latent variable (ability estimate-θ) estimated by IRT models is used as a matching variable for subgroups rather than observed scores. There are many different IRT based methods to detect DIF items, some of which are test of b difference, Lord's chi-square, Raju's area method, likelihood ratio test (LRT method), and item drift method. The Lord's chi-square DIF method (Lord, 1980) is an extension of the test of b difference and takes the other parameters into account. The major shortcoming of Raju's area method is that the exact areas between ICCs are infinite when guessing parameters are not equal (Hunter, 2014). In this project, Lord's chi-square DIF method was used to detect DIF items, because it takes more than one parameter into account when calculating DIF statistics and capable of detecting both uniform and non-uniform DIF in the presence of large sample size. The "difR" package installed in the R program was used to run DIF analyses. The formula for Lord's chi-square DIF methods is as follows:

$$Q_j = \left(v_{jR} - v_{jF}\right)' \left(\sum jR - \sum jF\right)^{-1} \left(v_{jR} - v_{jF}\right) \quad (1)$$

where $V_{jR} = (a_{jR}, b_{jR}, c_{jR})$ and $V_{jF} = (a_{jF}, b_{jF}, c_{jF})$ are the vectors of item parameters related to the reference group and focal group, respectively. Besides, the variance-covariance matrices of reference and focal groups are denoted by $\sum jR$ and $\sum jF$, respectively. The Q_1-statistic has chi-square distribution and its degrees of freedom is equal to the number of estimated parameters (Camilli, 2006; Lord, 1980). Previously research show that DIF results obtained from Lord's chi-squared test and Raju's unsigned area method are highly correlated (Millsap & Everson, 1993; Shepard, Camilli, &Williams, 1985). The most important disadvantage of the Lord's chi-squared test is that it tends to reject the null hypothesis of no DIF even when the discrepancy between ICCs of sub-groups is small in the presence of a large sample size (Camilli & Shephard,1994; Wiberg, 2006). Thus, a more stringent criterion should be used in the presence of a large sample size.

1.3 Differential test functioning (DTF)

Differential test functioning (DTF) values correspond to the total amount of DIF for the entire test. Therefore, it is equal to the sum of item DIF statistics in a test (Donovan, Drasgow & Probst, 2000; Ellis & Mead, 2000; Hunter, 2014; Nandakumar, 1993). There are different methods to calculate DTF such as Raju's DFIT (Raju, van der Linden & Fleer, 1995), and Mantel-Haenszel/Liu-Agresti method (Penfield & Algina, 2006). Raju's DFIT estimates DTF through calculating the squared difference between test characteristic curves, while the Mantel-Haenszel/Liu-Agresti method is based on variance estimates and tend to have higher DTF rates than Raju's DFIT (Hunter, 2014). Penfield (2005, 2013) developed a program called DIFAS which enables us to calculate both DIF and the Mantel-Haenszel/Liu-Agresti statistics. In this study, the MH-LA method was used to evaluate DTF associated with the EPT tests.

The formula for MH-LA DTF method proposed by Camilli and Penfield (1997) is as follows:

$$\tau^2 = \frac{\sum_{i=1}^{I}(\widehat{\varphi}_1 - \widehat{\mu})^2 - \sum_{i=1}^{I} s_i^{\,2}}{I} \quad (2)$$

where "I" represents the number of items; $\widehat{\psi_i}$ denotes MH log-odds ratio statistics; μ represents mean and s_i^2 represent the error variance of ψ. Some studies report weighted τ^2 statistics along with τ^2 statistics. The formula for weighted τ^2 is as follows:

$$\tau^2 = \frac{\sum_{i=1}^{I} W_i^2(\widehat{\varphi}_1 - \widehat{\mu})^2 - \sum_{i=1}^{I} W_i}{\sum_{i=1}^{I} W_i} \quad (3)$$

where w_i is equal to s_i^{-2}.

1.4 The purpose of the study

This study aims at examining the presence of items that function differently for the entire test and across different sub-domains for gender in English Placement Tests (EPT) to ensure the fairness of the test. For this purpose, item response theory (IRT) based Differential Item Functioning (DIF) at the item level and Differential Test Functioning (DTF) methods at the test level were implemented, respectively, to examine whether items function differently across different subgroups.

Considering the findings of previously conducted studies, this study aims to test five different hypotheses. The first hypothesis is that the factorial structure of the EPT remains unchanged across gender groups. The second hypothesis is that some of the EPT items are likely to exhibit DIF across gender. The third hypothesis assumes the existence of content specific DIF items at the item level. The fourth and fifth hypotheses are that the existence of DIF items affects the test scores for the entire test and each subdomain.

1.5 Research questions

1. Do the factorial structure of EPT for the entire test and each gender group support the unidimensionaliy?
2. Do items of the EPT function differently across gender (female vs. male)?
3. What is the distribution of DIF items across sub-domains (Reading Comprehension, Structure, and Compositional Analysis), when each domain is treated as a separate test?
4. Do test scores of the EPT exhibit differential test functioning (DTF) across gender (Female vs. Male)?
5. Do test scores of the EPT exhibit differential test functioning (DTF) across gender, when each domain is treated as a separate test?

2. Materials and Methods

2.1 The instrument and data

The EPT consist of three sub-domains that are reading comprehension, structure, and compositional analysis, respectively. It consists of 85 dichotomously scored items in which 22 items are related to reading comprehension, 43 items are related to structure and 20 items are related to compositional analysis, respectively. After the preliminary IRT-based item analyses, 2 items related to the reading comprehension domain, and 3 items related to the structure domain that showed misfit to the test were excluded. The final version of the EPT consists of 80 dichotomously scored items in which 20 items are related to reading comprehension, 40 items are related to structure and 20 items are related to compositional analysis, respectively.

The data for this study come from EPT 0105 test forms which were administered to 11,362 high-school graduates including 5665 females (49.85%) and 5,697 males (50.15%) in 2017. A relatively small sample data with 1000 cases were randomly drawn from the population and used to conduct the DIF and DTF analyses. The sample data comprise of 506 females (50.6%) and 494 (49.4%) males, respectively. The reason behind using the relatively small sample size is that the chi-square statistics are affected by the large sample size that increases the probability of committing Type-I error. In other words, some non-DIF items might be flagged as DIF items when the sample size is large.

2.2. Statistical analysis

This study used the quantitative descriptive research design to investigate the structure of the test, to detect the items that function differently across gender groups and the effect of these items on test scores at the test level. The first research question requires examining the factorial structure of EPT data. It is also necessary to see whether the assumption of unidimensionality is met since the IRT based DIF method will be implemented. A test is said to be unidimensional when there is one dominant factor (or latent variable) that underlies the scores obtained from the test. Thus, a one-factor CFA model was tested and fit measures of this one-factor CFA model were compared to see if the one-factor model fits the data. Besides, the one-factor CFA model was tested for both males and females to see whether the factorial structure remained the same across gender. A combination of data fit measures (goodness of fit statistics) such as the chi-square statistics, CFI (the comparative fit index), TLI (the Tucker-Lewis index) and RMSEA (root mean square error of approximation) for CFA models provides insight into the degree of data fit for the pre-specified model.

After checking the assumption of the IRT model, Lord's chi-square DIF method was used to detect items that exhibit DIF. The more stringent criterion for detecting DIF was favored and DIF analyses were employed to the sample data. Thus, the significance level of 0.01 (α=0.01) was used (rather than 0.05) with the detection threshold equal to 9.210. Along with Lord's chi-square DIF method, the Mantel-Haenszel/Liu-Agresti differential test functioning (DTF) method was employed to test the effects of DIF items at the test scores that might lead to unfair assessment. Penfield (2013) has suggested a set of criteria to assess the

degree of DTF for the Mantel -Haenszel/Liu-Agresti DTF method. Since it is based on the variance of DIF items, DTF statistics (t2) smaller than 0.07 are considered to be negligibly small, while DTF values (t2) between 0.07 and 0.14 indicate medium DTF effect and DTF values larger than 0.14 indicate large DTF effect, respectively. Hunter (2014) claims that the Mantel -Haenszel/Liu-Agresti DTF method is more stringent in general and shows higher rates of DTF compared to Raju's DFIT method. Therefore, DTF statistics larger than 0.14 is adopted as an indicator of substantial DTF for the test.

3. Results

In this section, firstly, CFA results that indicate the unidimensionality of each test are provided. Additionally, descriptive statistics and reliability coefficients of entire EPT tests and each subdomain of these tests are presented. Secondly, the results of the IRT-based Lord's Chi-square DIF method used to determine the items that function differently across gender for the entire test and each domain, are presented. Finally, the results of the Mantel -Haenszel/Liu-Agresti differential test functioning (DTF) method used to examine DIF effects across gender at the test level are provided in the following sections, respectively.

CFA results of the one-factor model and three-factor model, where each domain is treated as a factor, for the entire EPT, and each gender category are presented in Table 1.

Table 1. CFA results of the One-Factor CFA model of EPT Data Across Gender

Models	*Group*	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*90% for RMSEA*	
							LL	UL
One-factor Model	ALL	19376.905	3080	0.964	0.963	0.30	0.30	0.31
	Females	10385.585	3080	0.969	0.968	0.28	0.28	0.29
	Males	9731.157	3080	0.961	0.960	0.28	0.27	0.29

According to the goodness of fit statistics given in Table 1, both CFI and TLI statistics are above 0.95 indicating a good fit between model and data as Hu and Bentler (1999) suggested for both one-factor. Besides, the RMSEA values for the whole data set and each gender group are below the 0.06 criterion, and the 95% confidence interval of RMSEA is also below 0.06 indicating a good fit for both factors. However, chi-square values are statistically significant which are expected to be not statistically significant. The main reason behind this significant result might be the large sample size since the chi-square test results tend to be significant as the sample size increases. These results indicate that the one-factor CFA model shows a good fit to the data. Therefore, the EPT can be considered as unidimensional where all items load on one factor.

Table 2 presents descriptive statistics and reliability coefficients associated with the entire EPT test and with each subdomain. Two different reliability

coefficients, that are Cronbach's reliability coefficient and composite reliability coefficients (or latent variable modeling -LVM based reliability) were calculated. Composite reliability is calculated with factor loadings obtained from CFA, provided that the test is unidimensional. It provides more accurate reliability coefficients if essentially tau-equivalence does not hold and tends to yield higher reliability results when this assumption is violated.

Table 2. CFA results of the One-Factor CFA model of EPT Data Across Gender

Test/ Domain	*Mean*	*SD*	*r*	*Crombach- a*	*a-LL*	*a-UL*
EPT-ALL	49.38	17.68	0.96	0.95	0.95	0.95
Reading Comprehension	12.27	4.51	0.852	0.83	0.82	0.83
Structure	25.07	9.5	0.932	0.92	0.93	0.93
Compositional Analysis	12.04	4.46	0.844	0.83	0.84	0.84

According to results in Table 2, Cronbach – coefficients are substantially high with 0.95 for the entire test, 0.83 for reading comprehension (RC), 0.92 for structure, and 0.83 for compositional analysis. On the other hand, composite score reliability coefficients are somewhat higher compared to Cronbach coefficients. However, the discrepancy between composite reliability and Cronbach is negligible small indicating that the assumption of essentially tau-equivalence holds for EPT. Additionally, yielding substantially high-reliability coefficients is also an indicator of unidimensionality.

3.1 DIF and DTF results of the EPT

Table 3 provides the DIF results of the entire test regardless of different domains. The first column presents item numbers along with "rc", "st" and "ca" abbreviations that stand for each subdomain that are reading comprehension, structure and compositional analysis, respectively. The second and third column presents Lord's chi-square DIF statistics and corresponding significance test results (p-values) for each item.

Table 3. DIF statistics of all items in EPT test 0105

Item no	*Statistics*	*p-value*	*Item no*	*Statistics*	*p-value*
rc1	0.983	0.612	st21	4.383	0.112
rc2	0.030	0.985	st22	1.389	0.499
rc3	1.454	0.483	st23	4.193	0.123
rc4	**30.521**	**0.000**	**st24**	**12.239**	**0.002**
rc5	2.054	0.358	st25	0.778	0.678
rc6	0.036	0.982	st26	1.482	0.477
rc7	**22.769**	**0.000**	st27	1.487	0.476
rc8	7.114	0.029	st28	2.788	0.248
rc9	2.710	0.258	st29	1.666	0.435

rc10	4.799	0.091	st30	0.345	0.841
rc11	7.335	0.026	st31	0.241	0.887
rc12	7.093	0.029	st32	1.479	0.477
rc13	2.741	0.254	st33	5.823	0.054
rc14	2.736	0.255	st34	0.681	0.712
rc15	0.330	0.848	st35	6.295	0.043
rc16	4.843	0.089	st36	5.957	0.051
rc17	2.439	0.295	st37	0.706	0.703
rc18	3.727	0.155	st38	8.474	0.015
rc19	1.034	0.597	**st39**	**11.021**	**0.004**
rc20	5.446	0.066	st40	2.700	0.259
st1	0.517	0.772	st41	1.671	0.434
st2	7.003	0.030	ca1	3.237	0.198
st3	**12.894**	**0.002**	ca2	0.041	0.980
st4	1.212	0.545	ca3	3.856	0.145
st5	2.028	0.363	ca4	0.593	0.744
st6	4.257	0.119	ca5	3.949	0.139
st7	1.478	0.478	ca6	4.428	0.109
st8	0.129	0.938	ca7	0.816	0.665
st9	0.065	0.968	ca8	1.269	0.530
st10	5.662	0.059	ca9	1.066	0.587
st11	1.397	0.497	ca10	2.331	0.312
st12	0.720	0.698	ca11	2.348	0.309
st13	5.954	0.051	ca12	5.999	0.050
st14	2.261	0.323	ca13	3.566	0.168
st15	2.651	0.266	ca14	7.381	0.025
st16	**11.778**	**0.003**	ca15	1.725	0.422
st17	0.338	0.845	ca16	0.566	0.754
st18	2.831	0.243	ca17	1.043	0.594
st19	4.721	0.094	ca18	1.059	0.589
st20	0.054	0.974	ca19	7.728	0.021

According to results in Table 3, six items (rc4, rc7, st3, st16, st24, st39) out of 80 items of EPT had Lord's chi-square statistics greater than DIF detection threshold (9.21) and are detected as DIF items. Although chi-square statistics associated with rc4 and rc7 are substantially high, the other four items' chi-square statistics are around 12 and are close to the DIF detection threshold. DIF results also indicate that 2 out 6 DIF items are associated with reading comprehension, while 4 out 6 DIF items are associated with the structure domain. None of the items of the compositional analysis domain are detected as DIF. Figure 1 depicts item characteristic curves (ICC) of focal (male) and reference (female) groups for each item detected as DIF. The straight-line represents ICC associated with the focal group, while the dotted line represents ICC associated with the reference group. The lines in the ICCs represent the probability of answering an item correctly across the ability range (θ) for each gender group. The discrepancy between the lines indicates the existence and the amount of DIF effect.

One can observe from ICCs given in Figure 1 that item 4 (rc4) and item 23 (st3) exhibit uniform DIF meaning that the discrepancy of ICCs between males and females is consistent across the entire range of abilities (Hambleton et al., 1993).

Moreover, item 4 shows DIF favoring male students, while item 23 shows DIF favoring female students. On the other hand, the other 4 items exhibit non-uniform DIF indicating that discrepancies between ICCs of DIF items are not consistent across the ability distribution. Moreover, one can observe that male students perform better at low ability levels, while female students perform better at high ability level for each non-uniform DIF items.

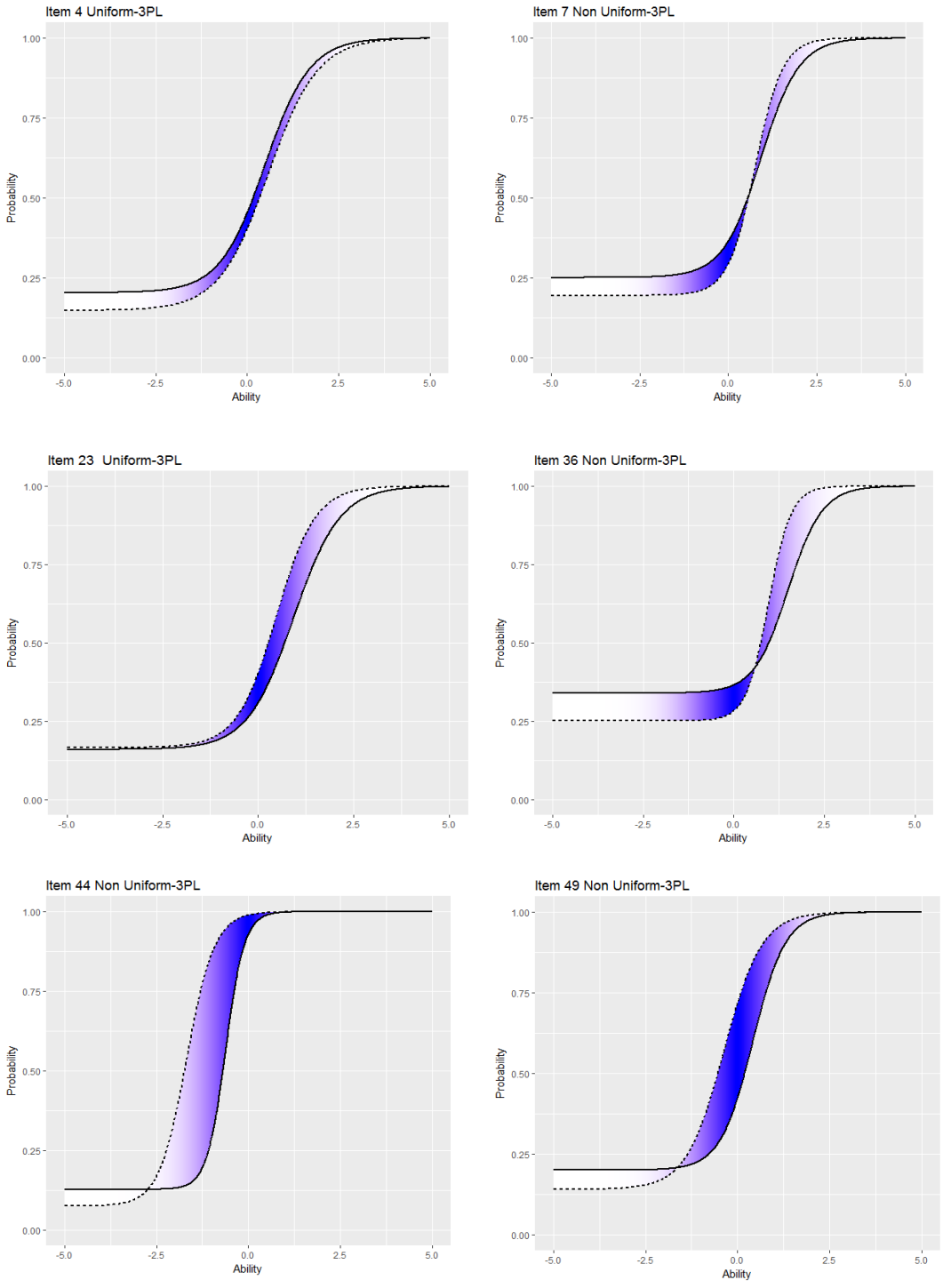

Figure 1. Item Characteristic Curves for DIF items of the EPT

3.2 DIF results across sub-domains

Table 4 provides DIF results of each subdomain: reading comprehension, structure, and compositional analysis. The first 3 columns present item no, Lord's chi-square DIF statistics, and corresponding significance test results (*p*-values) related to reading comprehension. DIF results of structure are followed by reading comprehension and compositional analysis domains.

Table 4. DIF statistics across sub-domains of EPT

Structure			*Reading comprehension*			*Compositional analysis*		
Item no	*Statistics*	*p-value*	*Item no*	*Statistics*	*p-value*	*Item no*	*Statistics*	*p-value*
st1	0.542	0.763	rc1	1.249	0.536	ca1	4.305	0.116
st2	7.942	0.019	rc2	0.089	0.956	ca2	0.558	0.757
st3	**11.870**	**0.003**	rc3	1.559	0.459	ca3	2.700	0.259
st4	0.346	0.841	**rc4**	**24.107**	**0.000**	ca4	1.110	0.574
st5	1.950	0.377	rc5	0.762	0.683	ca5	3.525	0.172
st6	4.641	0.098	rc6	0.158	0.924	ca6	4.757	0.093
st7	1.052	0.591	**rc7**	**16.801**	**0.000**	ca7	2.479	0.290
st8	0.317	0.854	rc8	6.881	0.032	ca8	0.267	0.875
st9	0.162	0.922	rc9	2.559	0.278	ca9	2.063	0.357
st10	4.603	0.100	rc10	3.028	0.220	ca10	4.878	0.087
st11	0.700	0.705	rc11	5.228	0.073	ca11	0.549	0.760
st12	0.419	0.811	rc12	2.969	0.227	ca12	6.212	0.045
st13	7.077	0.029	rc13	5.676	0.059	ca13	3.670	0.160
st14	2.632	0.268	rc14	2.039	0.361	ca14	2.390	0.303
st15	3.283	0.194	rc15	0.134	0.935	ca15	2.725	0.256
st16	**11.177**	**0.004**	rc16	2.469	0.291	ca16	0.747	0.689
st17	0.050	0.975	rc17	1.526	0.466	ca17	2.577	0.276
st18	3.760	0.153	rc18	0.690	0.708	ca18	0.274	0.872
st19	3.601	0.165	rc19	2.021	0.364	ca19	6.162	0.046
st20	0.000	1.000	rc20	4.331	0.115			
st21	3.772	0.152						
st22	1.390	0.499						
st23	5.243	0.073						
st24	**13.468**	**0.001**						
st25	0.770	0.681						
st26	1.505	0.471						
st27	0.247	0.884						
st28	0.941	0.625						
st29	0.652	0.722						
st30	0.920	0.631						
st31	0.433	0.805						
st32	1.498	0.473						
st33	3.856	0.145						
st34	0.390	0.823						
st35	7.801	0.020						
st36	5.763	0.056						
st37	0.779	0.678						
st38	8.422	0.015						
st39	9.671	0.008						
st40	3.021	0.221						
st41	0.970	0.616						

DIF results in Table 4 indicate that 2 items (rc4, rc7) in reading comprehension and 4 items (st3, st16, st24, st39) in structure domains are detected as DIF. Additionally, none of the items of the compositional analysis domain are detected as DIF. One can notice that those same items are detected as DIF items for the entire test and each subdomain. Moreover, chi-square statistics associated with each DIF item across sub-domains tend to decrease somewhat compared to DIF results of the entire test in Table 3. Especially, decrements in DIF statistics are quite obvious for rc4 and rc7 and the DIF statistic of st39 (9,671) is close to the DIF detection threshold (9.21).

3.3 Differential test functioning (DTF) Results

In this study, the Mantel-Haenszel/Liu-Agresti differential test functioning (DTF) method (Penfield, 2013) which is based on variance estimates of DIF items, was used to examine DIF at test level. Table 5 provides DTF statistics including variance estimates (t2), weighted variance estimates (Weighted t2), associated standard errors (SE), and z-scores for each DTF statistic for the entire test and each subdomain.

Table 5. DTF results for the entire test and each subdomain

Test/domain	*Statistic*	*Value*	*SE*	*Z*
EPT-All	t^2	0.068	0.012	5.667
	Weighted t^2	0.06	0.01	6.000
Reading Comprehension	t^2	**0.097**	0.032	3.031
	Weighted t^2	**0.072**	0.024	3.000
Structure	t^2	0.067	0.016	4.188
	Weighted t^2	0.06	0.015	4.000
Compositional Analysis	t^2	0.032	0.012	2.667
	Weighted t^2	0.03	0.011	2.727

According to results in Table 5, the DTF variance associated with the entire test (0.068) is less than 0.07 indicating that the DTF effect of EPT is negligibly small. Moreover, this indicates that test scores do not function differently across gender at test level. Although 6 items detected as showing DIF, DTF results indicate that DIF effect cancels each out at test level, because some of them show DIF in favor of males, while some of them are in favor of females. When it comes to DTF variance associated with sub-domains, structure and compositional analysis domains yield DTF variance less than 0.07 indicating that DTF statistics associated with these domains are negligibly small. Moreover, DTF associated with reading comprehension (0.097) falls within 0.07 and 0.14 indicating a moderate DTF effect. However, the weighted variance associated with the reading comprehension domain (0.072) is close to 0.07 and can be considered as negligible small. The DTF variance of the compositional analysis domain (0.032) is relatively small compared to the other two domains since one item is detected as DIF supporting the DIF results at the item level. Thus, both negligible small DTF effects of the entire test and each domain indicate that DIF effects cancel each other at the test level. 4.

4. Discussion

In this study, the IRT-based Lord's Chi-square DIF method was utilized to determine the items functioning differently in the English Placement Test (EPT) across gender for the entire test and each subdomain. Moreover, the Mantel - Haenszel/Liu-Agresti (MH-LA) differential test functioning (DTF) method was used to examine the DIF effect at the test level. The results of DIF and DTF analyses for the EPT were evaluated and compared at the item and test level.

DIF analysis results indicate that 6 items (rc4, rc7, st3, st16, st24, st39) in EPT exhibits DIF regardless of test domains. When it comes to the distribution of DIF items across sub-domains, two DIF items are associated with the reading comprehension domain and the rest are associated with the structure domain. Moreover, none of the items of the compositional analysis domain is detected as exhibiting DIF. The DIF results across sub-domains, where each subdomain is treated as an independent test, yield parallel results with the entire test. Moreover, the number of DIF items, items detected as DIF, and the distribution of DIF items across sub-domains are identical with the entire test. However, the chi-square statistics associated with each DIF item across sub-domains tend to decrease somewhat when compared to the DIF results of the entire test. These results signal the existence of content specific DIF effect for the entire test. In other words, some domains, such as reading comprehension and structure, appear to be more prone to the DIF. These content-specific DIF effects might occur due to unintended latent traits (Ercikan et al., 2010) item contents such as cultural background or item properties. These unintended content-related factors increase the likelihood of occurrence of DIF (Martinkova et al., 2017).

Item characteristic curves (ICCs) related to DIF items given in Figure 1 for focal (male) and reference (female) groups provide information about the type of DIF (uniform or non-uniform DIF) and behavior of items across ability levels. The ICCs associated with each gender group reveal that two items (item 4 and item 23) exhibit uniform DIF. For these two DIF items, the male students perform better than female students on item 4, while female students show better performance on 23 compared to male students. Moreover, the other 4 items of EPT exhibit non-uniform DIF indicating that discrepancy between ICCs of DIF items are not consistent across the ability distribution. For these non-uniform DIF items, male students perform better than female students at low ability levels, while female students perform better than male students at high ability levels for each non-uniform DIF items. These types of items require revision of content experts to define the source of DIF and to decrease the unfair effects of DIF on the evaluation process in large scale assessments (Penfield & Lee, 2010; Martinkova et al., 2017).

Differential functioning at item level and test level appear to be associated and DTF is considered to be the sum of DIF for compensatory DIF defined by Raju and his colleague (Raju & Ellis, 2003). DTF results for the entire test of EPT show that the DIF variance associated with the entire test is less than 0.07 indicating that the DTF effect of EPT is negligibly small. Although 6 items detected as showing DIF, DTF results indicate that DIF effect cancels each out at test level,

because for some of them females outperform males, while males outperform females for the others. For compensatory DIF, there is a cancellation effect in which the DIF effect may cancel each out in the presence of items favoring different subgroups at test level (Flora, Curran, Hussong, & Edwards, 2008; Hunter, 2014; Nandakumar, 1993; Takala & Kaftandjieva, 2000). These results assure that EPT test scores does not function differently across gender and supports the fairness and validity of the test results at the test level.

When it comes to DTF effects across sub-domains, structure, and compositional analysis domains have DTF variance less than 0.07 indicating that DTF effects associated with these domains are negligibly small. However, DTF associated with reading comprehension falls within 0.07 and 0.14 indicating moderate DTF effect, while weighted variance associated with reading comprehension domain is close to 0.07 and could be considered as negligible small. The relatively larger DTF effect associated with the reading comprehension domain might be an indicator of the existence of a construct-irrelevant latent factor such as the degree of vocabulary knowledge of test takers that have a benign effect on test results (Jang & Roussos, 2009). Moreover, the relatively larger DTF effects associated with reading comprehension and structure domains reveal that the existence of DIF effects at item level influences the DTF results. These results might also imply the existence of content specific DTF effect.

Chubbuck and his colleagues (2016) examined the performance of gender groups on sentence-completion and reading comprehension questions using the Mantel–Haenszel and standardized DIF methods. They found out the content specific DIF in sentence-completion items in which males outperformed females in reading comprehension items (Wedman, 2018). The findings of the aforementioned studies support the results of this study concerning the occurrence of content specific DIF. Another factor that might cause DIF is the language skills of non-native test takers that take a test in a language other than their mother tongue. The deficiency in their language skill or failure in wording the content clearly in the item might lead to DIF between sub-groups (Siegel, 2007, Wedman, 2018). The results of DIF and DTF induce item bias and violation of test fairness when a large number of items are in favor of a certain group and when unintended construct irrelevant factors are defined as a source of DIF (Zhu & Aryadoust, 2020). The relatively small number of DIF items and negligibly small DTF effects of the entire test indicate that the fairness of test scores is achieved for the EPT. However, it is of great importance to use methods such as DIF and DTF to examine the fairness of the test across gender groups and to ensure equality between males and females. On the other hand, the results of this study showed that unintended factors, such contents favoring a certain group, might lead to the DIF effects which can only be controlled by content experts. Thus, another way of ensuring test fairness requires selecting the contents that are relevant to each gender group.

5. Conclusion

DIF analysis is one of the most important methods employed to ensure the validity of the test and fairness of test score interpretation (Zumbo, 2007). The First step in DIF is to use statistical methods to determine DIF items. This step is followed by deciding whether to remove or to revise these items since statistically significant DIF results do not always indicate biased items. It requires a comparison of differential functioning results at item and test level and involvement of content experts for the final decision. There are different approaches to deal with items detected as DIF. Some researchers suggest removing DIF items to reduce DTF effect (Raju et al., 1995) while others suggest consulting test developers and content experts to examine the structure of test and items before removing DIF items and try to determine what exactly caused differential functioning (Martinkova et al., 2017; Penfield & Lee, 2010). Therefore, items with substantially high DIF values (rc4 and rc7 items) should be examined by content experts. Because, removing DIF items without any evaluation does not ensure the fair test (Clauser & Mazor,1998; Gierl et al., 2001; Hunter, 2014), specifically, when DTF effects of test forms are negligibly small and DIF effects cancel each out at test level.

Some researchers who claim that removing DIF items may lead to weaker tests (rather than fair test) regarding the representation of constructs and variance explained by these items (Roznowski & Reith 1999). Therefore, consulting with test developers and content experts before removing the DIF items is suggested. It is also suggested to investigate the effects of other potential factors on DIF such as item order and mother tongue effects along with unintended content specific factors to explain DIF effect in the context of language testing. It is acknowledged that detecting DIF items might require using a combination of DIF methods to increase the accuracy of the results. This study is limited to detecting items that function differently across the gender groups for each content domain. The existence of DIF across other subgroups, such as native vs non-native speakers and across nationalities could be studied. Although this study provided evidence about the existence of content-specific DIF effect as a potential source of DIF, it was not possible to examine the content of each DIF item with content experts since the EPT items were not released. Another limitation is that the unidimensionality of test was addressed with the first research question, while the effect of the multidimensionality and the existence of unintended latent factor on DIF and DTF results were not taken into account.

6. References

Anastasi, A., & Urbina, S. (1996). *Psychological Testing*. Upper Saddle River, NJ: Prentice Hall.

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing*, *33*(4), 529-553. doi:10.1177/0265532215594640

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), 176-181. doi:10.1097/01.mlr.0000245143.08679.cc

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 221-256). Westport: American Council on Education & Praeger Publishers.

Camilli, G., & Penfield, D. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*(2), 123-139. doi:10.1111/j.1745-3984.1997.tb00510.x

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

Chubbuck, K., Curley, W. E., & King, T. C. (2016). *Who's on first? Gender differences in performance on the SAT test on critical reading items with sports and science content* (Report No. RR-16-26). Princeton, NJ: Educational Testing Service.

Clauser, B., & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44. doi:10.1111/j.1745-3992.1998.tb00619.x

Donovan, M. A., Drasgow, F., & Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology, 85*(2), 305-313. doi:10.1037/0021-9010.85.2.305

Drabinová, A., & Martinková, P. (2016). Detection of differential item functioning with non-linear regression: Non-IRT approach accounting for guessing. Retrieved from http://hdl.handle.net/11104/0259498

Ellis, B. B., & Mead, A. D. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement, 60*(5), 787-807. doi:10.1177/00131640021970781

Ellis, B., & Raju, N. (2003). Test and item bias: What they are, what they aren't, and how to measure them. In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators.* (pp. 89-98). Greensboro, N.C.: CAPS.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon F., & Lacroix S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24-35. doi:10.1111/j.1745-3992.2010.00173.x

Education & Training Evaluation Commission. (2020). *Language Test*. Retrieved from https://etec.gov.sa/en/productsandservices/Qiyas/lingual/Pages/default.aspx

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25*(3), 283-291. doi:10.7334/psicothema2013.97

Federer, M. R., Nehm, R. H., & Pearl, D. K. (2016). Examining gender differences in written assessment tasks in biology: a case study of evolutionary explanations. *CBE – Life Sciences Education*, *15*(1), ar2. doi:10.1187/cbe.14-01-0018

Ferne, T., & Rupp, A. A. (2007). A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations. *Language Assessment Quarterly*, *4*(2), 113–148. doi:10.1080/15434300701375923

Flora, D., Curran, P., Hussong, A., & Edwards, M. (2008). Incorporating measurement nonequivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling, 15*(4), 676-704. doi:10.1080/10705510802339080

Gierl, M., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36. doi:10.1111/j.1745-3992.2001.tb00060.x

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11), 182-188. doi:10.1097/01.mlr.0000245443.86671.c4

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*, 1-18. Retrieved from https://eric.ed.gov/?id=ED356264

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313-34. doi:10.1207/s15324818ame0204_4

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park. Calif.: Sage Publications.

Hernández, A., Tomás, I., Ferreres, A., & Lloret, S. (2015). Tercera evaluación de test editados en España [Third evaluation of tests published in Spain]. *Papeles del Psicólogo, 36*(1), 1-8. Retrieved from http://www.papelesdelpsicologo.es/pdf/2484.pdf

Hope, D., Adamson, K., McManus, I. C., Chris, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC Medical Education, 18,* 1-7. doi:10.1186/s12909-018-1143-0

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi:10.1080/10705519909540118

Hunter, C. (2014). A simulation study comparing two methods of evaluating differential test functioning (DTF): DFIT and the Mantel-Haenszel/Liu-Agresti variance (Doctoral Dissertation). Georgia State University, Atlanta, GA, United States.

Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing*, *9*(3), 238–259. doi:15305050903107022.

Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the Health Professions, 28*(3), 283-294. doi:10.1177/0163278705278276

Lin, J., & Wu, F. (2003). *Differential performance by gender in foreign language testing* [Poster presentation]. The annual meeting of the National Council on Measurement in Education, Chicago, United States.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luo, Y., & Al-Harbi, K. (2016). The Utility of the bifactor method for unidimensionality assessment when other methods disagree: An empirical illustration. *Sage Open, 6*(4), 1-7. doi:10.1177/2158244016674513

Magis, D., & Facon, B. (2012). Angoff's Delta method revisited: improving the DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology, 65*(2), 302-321. doi:10.1111/j.2044-8317.2011.02025.x

Martinková, P., Drabinová, A., Liaw, Y., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *CBE—Life Sciences Education*, *16*(2), 1-13. doi:10.1187/cbe.16-10-0307

Millsap, R. E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Medical Care*, *44*(11), 171-175. doi:10.1097/01.mlr.0000245441.76388.ff

Millsap, R., & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334. doi:10.1177/014662169301700401

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293-311. doi:10.1111/j.1745-3984.1993.tb00428.x

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

Pae, T., & Park, G. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475-496.

Penfield, R. D. (2005). DIFAS: Differential item functions analysis system. *Applied Psychological Measurement*, *29*(2), 150-151. doi:10.1177/0146621603260686

Penfield, R. (2013). DIFAS 5.0: Differential item functions analysis system. User's manual. Retrieved from https://soe.uncg.edu/wp-content/uploads/2015/12/DIFASManual_V5.pdf

Penfield, R., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*(4), 295-312. doi:10.1111/j.1745-3984.2006.00018.x

Penfield, R., & Lee, O. (2010). Test-based accountability: potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching*, *47*(1), 6-24. doi:10.1002/tea.20307

Raju, N., & Ellis, B. (2003). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behaviour in organizations: Advances in measurement and data analysis.* (pp. 156-188). San Francisco: Jossey-Bass.

Raju, N., van der Linden, W., & Fleer, P. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353-368. doi:10.1177/014662169501900405

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*(2), 248-269. doi:10.1177/00131649921969839

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194. doi:10.1007/BF02294572

Shepard, L. A., Camilli, G., & Williams, A. F. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*(2), 77-105. doi:10.1111/j.1745-3984.1985.tb01050.x

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching*, *44*(6), 864–881. doi:10.1002/tea.20176

Stage, C. (2005). *Socialgruppsskillnader i resultat på högskoleprovet [Social group differences in scores on the Swedish Scholastic Assessment Test]*. (Report No. BVM 11:2005). Umeå: Umeå University.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 117*(3), 323-340. doi:10.1191/026553200678030346

Wedman, J. (2018). Reasons for Gender-Related Differential Item Functioning in a College Admissions Test. *Scandinavian Journal of Educational Research, 62*(6), 959-970, doi:10.1080/00313831.2017.1402365

Wiberg, M. (2006). Gender differences in the Swedish driving-license test. *Journal of Safety Research, 37*(3), 285-291. doi:10.1016/j.jsr.2006.02.005

Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*, *33*, 1-24. doi:10.1080/09588221.2019.17047884

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. doi:10.1191/0265532203lt248oa

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. doi:10.1080/15434300701375832