THE JOURNAL OF ASIA TEFL
Vol. 16, No. 2, Summer 2019, 674-682
http://dx.doi.org/10.18823/asiatefl.2019.16.2.15.674

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

A Gender–Related Differential Item Functioning Study of an English Test

Masoud Geramipour
Kharazmi University, Iran

Niloufar Shahmirzadi
Islamic Azad University, Central Tehran Branch, Iran

Introduction

Standardized tests are widely applied in many settings, including education, psychology, business, and medicine. In this case, numerous disciplines have been used to identify the potential source of invariability in gender, age, cultural background and so on in research (Andersen, 1967). To delve into the issue, one of these sources that is to say gender has detected in Differential Item Functioning (DIF) as systematic measurement variability which leads to a number of problems including errors in hypothesis testing, population forecasts, policy planning and implementation, and misguided research on disparities (Perkins, Stump, Monahan, & McHorney, 2006).

In language testing when two groups of test takers with the same ability level cannot respond equally well to an item, the issue of DIF may occur (Clauser & Mazor, 1998). In this case if a group does not have an equal chance of correctly answering an item, it may be considered a biased item. Zumbo (1999) proposes that DIF can be applied to detect bias at the item level. A number of studies have also been conducted on factors such as gender (Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000), language background (Brown, 1999; Chen & Henning, 1985; Elder, 1996; Kim, 2001), and academic background or content knowledge (Alderson & Urquhart, 1985; Hale, 1988; Pae, 2004). Camilli (2006) and Wiberg (2007) claim that DIF would be synonymous with statistical bias where one or more parameters of the statistical model are over/under-estimated. Therefore, whenever an item is identified as a DIF item, it would be worth investigating whether the item is a case of bias. Such bias can be due to construct–irrelevant variance.

DIF study is conducted with two groups called a focal and reference group. The former group could refer to a minority test taker group and the latter group pertains to those potentially advantaged by the test. Typically, DIF is attributed to multidimensionality in which the primary dimension(s) is the focus of the test and the additional attribute which may be a nuisance dimension may cause DIF (Shealy & Stout, 1993; Roussos & Stout, 1996). Practically, some test items are intentionally developed to measure multiple traits (Clauser, Nungester, Mazor, & Ripkey, 1996); thus, it is vital to detect multiple dimensions before seeking bias (Clauser et al., 1996; Mazor, Hambleton, & Clauser, 1998).

In addition, two types of DIF could be identified called uniform and non–uniform. In uniform DIF, test takers from one group perform better than test takers in another group on all ability levels. That is to say, all members of test takers could outperform other groups who are at the similar levels of ability. In non-uniform DIF, on the other hand, test takers from one group may perform better than test takers in another group but not on all levels of ability. Thus, there is an interaction between grouping and the ability levels.

Ordinal Logistic framework also has a number of advantages over other techniques such as flexibility and incorporating IRT ability estimates, determining the item parameters which are all led into declaring items to have DIF. The second advantage of ordinal logistic algorithm is speed. However, a vital limitation for ordinal LR for DIF detection is the necessity to be familiar with both IRT and LR. In ordinal LR techniques, having a sufficient sample size is crucial for the stable estimation of model parameters. In sum, DIF detection can heighten the evaluation in high speed. It also facilitates the use of IRT derived ability estimates and provides flexible assessment of DIF in test items.

Gender is a crucial factor in DIF detection. For example, Song, Cheng, and Klinger (2015) examined gender and academic background and concluded that motivation and different lifestyles may cause some changes in the process of learning and performance. Ogretmen (2015) reported that in advanced reading comprehension section of an English high-stakes Test DIF can be observed especially in some subskills. Thus, it is vital to detect gender bias in DIF detection.

This study was aimed at detecting and measuring DIF and DTF in one of the most vital national English tests for senior high school students in Iran. The Lordif package of the R program was used by the researchers so as to observe the performance of the hybrid ordinal logistic regression on a real partial credit data set. The following is the research question of the present study:

> Which items of the English test have uniform or non-uniform DIF using hybrid ordinal logistic regression?

Materials and Method

Participants

The data for the present research was collected through multistage cluster sampling from among 1000 participants who were doing their diploma studies majoring in Mathematics, Natural Sciences, and Human Sciences at the high school level in Iran. All the participants were between the age range of 17 to 18, including both females (52.8%) and males (47.2%) which were proportionately selected.

Data Collection

The instrument for the present study was a national test from the National English Exam of senior high school students (K-11) which is held each year in June. The results of exam are crucial for the test takers' future academic life. The test is generally designed and written by a panel of anonymous experts in English language; however, the quality of the test needs to be investigated. The test consists of 30 open-ended questions including reading, writing skills, vocabulary and grammar subskills. The total allotted time was 70 minutes. Regarding each question weight, it is worth noting that they carry Partial Credit Model (PCM) that is to say right response to each subsection of a question received credit.

Procedure

The corpus for the study was collected from among national senior high school students (K-11) who sat for the high-stakes test in all subjects including English. Thus, the result of this study is of paramount importance for both test takers and English language teachers. Because in DIF studies, it is generally

assumed that test takers from different groups with the same underlying ability have different probability of responding to the same test items which were compared through reference group (female) and focal group (male) here.

The researchers selected a sample of 1000 test takers' raw data into the R software. Then, it was analyzed by applying iterative hybrid ordinal logistic regression through the Lordif package (Choi, Gibbons, & Crane, 2016). In the following section, the results of the study are provided in detail.

Results

The present section presents the results of the English examination of senior high school participants. However, before doing the major DIF/DTF analysis, it was mandatory to investigate the unidimensionality assumption of the item response data. According to Angoff (1993) "all the methods and techniques that have been developed to identify DIF in the items assume that the group of items, or the test that contains the item, is homogeneous and unidimensional" (p. 14). Unidimensionality of the item response data was also investigated through the MIRT package of R before analyzing DIF. According to Li, Hunter and Lei (2015), some related techniques in absolute model fit measures for the present study are Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMSR) which is based on the average differences between observed and predicted correlation matrices along with TLI (Tucker-Lewis Index), and CFI (Comparative Fit Index). To check absolute model fit indices, General Partial Credit (GPCM) model was fitted to the response data. The fit indices of the model are shown in Table 1.

TABLE 1
Fit Indices

RMSEA	SRMSR	TLI	CFI
0.0517	0.0625	0.9757	0.9774

The fit indices of the unidimensional model revealed that the goodness of fit is tenable. Thus, the assumption of unidimensionality of the selected model for the test was confirmed. That is to say, RMSEA < .05 (.0517) was good, and SRMSR < .05 (.062) was also almost acceptable. Thus, these indices confirmed that the GPCM for estimating ability scores based on polytomous IRT model was valid enough to be combined with ordinal logistic regression to detect DIF/DTF as a hybrid method. Interestingly enough, TLI (Tucker-Lewis Index) > .95, and CFI (Comparative Fit Index) > .95 confirmed the good model fit since both scored approximate to cutoffs including .9757 and .9774, respectively.

In what follows, a graphical display of items is provided with regard to gender in Figures 1 to 5. Figure 1 depicts trait distributions of reference (female) and focal (male) groups which followed smoothed histograms of the reference group (solid line) and focal group (dashed line) of participants. Generally, attempts were made to find which group suffered from DIF in replying to test items. Thus, they cannot reveal participants' true ability. Figure 1 also indicates that the focal group moved dramatically up and down, but reference group followed a rather smoothed line before reaching the peak, and finally both reference and focal groups overlapped in the distributions.

Researchers, in addition, used the likelihood ratio (LR) x^2 test (criterion = x^2) as the detection criterion at the α level of .01, and McFaddn's pseudo R^2 (default) as the magnitude measure. In addition, the plot function in Lordif presents (see Figure 1) the theta distributions for the reference and focal groups. On average both reference and focal groups had non-uniform DIF which are provided in detail below.

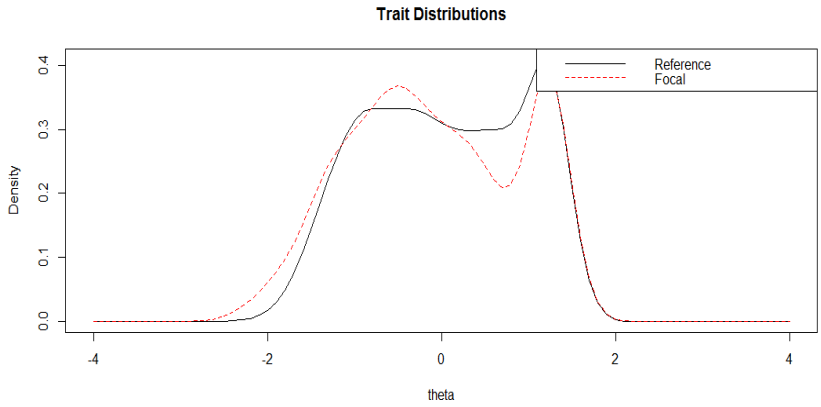

Figure 1. Distribution of reference and focal groups.

The top left plot in Figure 2 shows item true score functions–item 30 based on group specific item parameter estimates. The slope of the function for the reference group was slightly higher than that of the focal group, indicating non–uniform DIF. The LR x^2 test for uniform DIF, comparing Model 1 and Model 2, was not significant (p = .07), whereas the 1–df test for comparing Model 2 and Model 3 (p = .0013). It is interesting to note that having 2–df tests (comparing Models 1 and 3) were used as the criterion for female, this item had not outperformed at α = .01 (p = .00011).

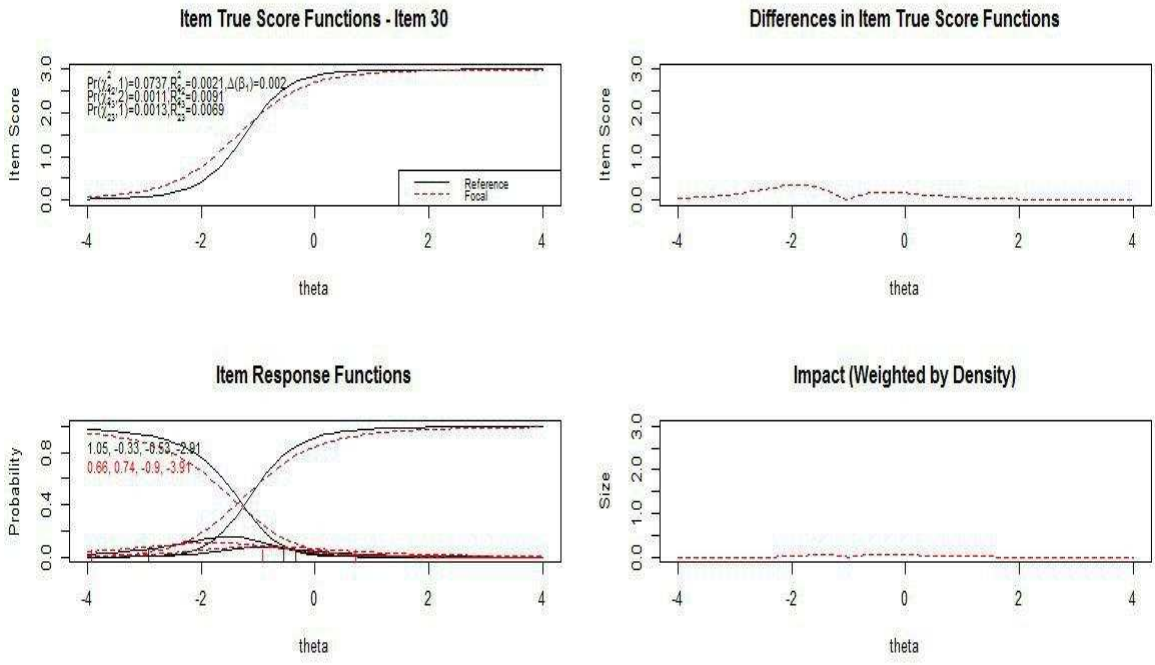

Figure 2. Item 30 plot.

The bottom plot on the left in Figure 2 juxtaposed the item response functions for reference and focal groups. The non–uniform component of DIF revealed by the LR x^2 test can also be observed in the difference of the slope parameter estimates (1.05 vs. 0.66). Although there was not a meaningful uniform DIF, the difference in the second category threshold values for the two groups were noticeable (-0.53 vs. –0.9). For polytomous items, a single item-level index of DIF may not provide adequate information concerning response categories (or score levels) contributing to the DIF effect. It is noteworthy to

mention that "the combination of visual and model–based approached in Lordif provided useful diagnostic information at the response category level, which can be systematically investigated under the differential step functioning framework" (Penfield, 2007, p. 44; Penfield, Gattamorta, & Childs, 2009, p. 28).

The top right plot in Figure 2 represents the expected impact of DIF on scores as the absolute difference between the item true score functions (Kim, Cohen, Alagoz, & Kim, 2007). There was a difference in the item true score functions peaking at approximately ø = – 1.5, but the density-weighted impact (depicted in the bottom right plot) was negligible because few subjects had the trait level in this population. When weighted by the focal group trait distribution, the expected impact became negligible which was apparent in the small McFadden's pseudo R^2 measures (printed on the top left plot) that are $R^2_{13} = 0.0091$ and $R^2_{13} = 0.0069$.

Figure 3 indicates the plots for item 29, which manifested statistically meaningful uniform DIF. The LR x^2_{13} was also significant; however, as the LR x^2_{23} was non-meaningful this result suggests the DIF was primarily uniform. In fact, the item response functions suggested that the uniform DIF was due to the first category threshold value for the focal group being smaller than the reference group (0.03 vs. 0.06).

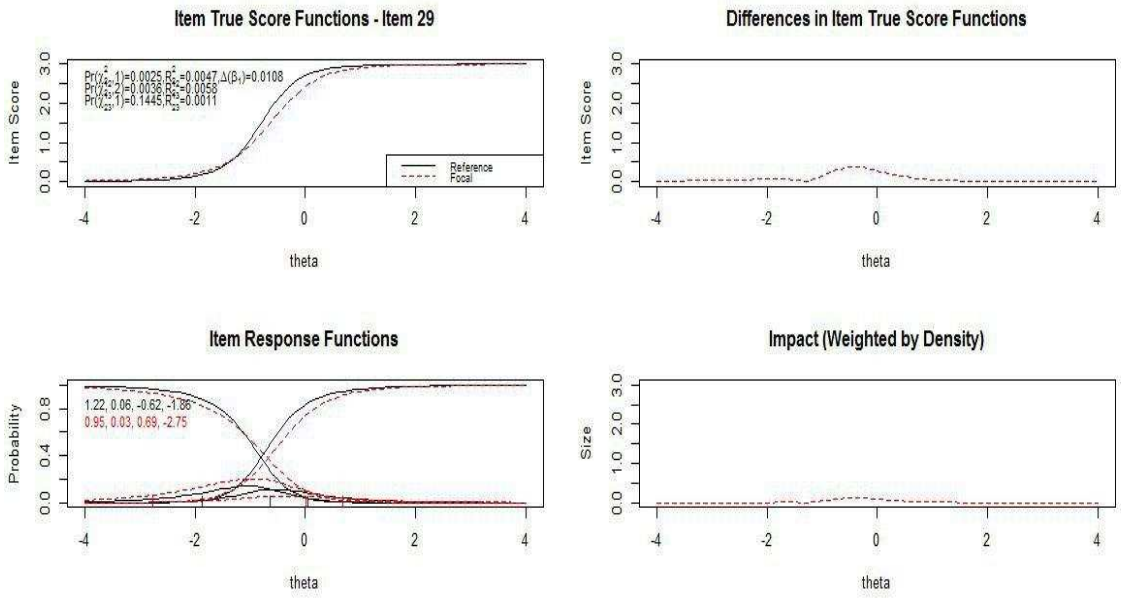

Figure 3. Item 29 plot.

Figure 4 also presents non–uniform DIF for item 28. In the meanwhile, x^2_{12} and x^2_{13} were not significant ($p < 0.001$) along with non-significant x^2_{23}. McFadden's R^2 also changed for uniform DIF, which carried a negligible effect size (Cohen, 1988). And, the item response functions showed that the category threshold parameters for the focal group were uniformly smaller than those for the reference group.

Figure 5, moreover, displays uniform DIF for 30 items since all items almost aggregated over all the items in the test (left plot) or over the subset of items found to have DIF (right plot), differences in item characteristic curves may also become negligibly small due to the cancelling of differences in opposite directions, which was what occurred here.

As mentioned earlier gender played a significant role in the study; thus, the number of categories was divided into two groups of male and female in an English test comprised 30 items which were scored based on the GPCM model. In so doing, it is shown that only *three* items had significant DIF, and the obtained effect size was negligible.

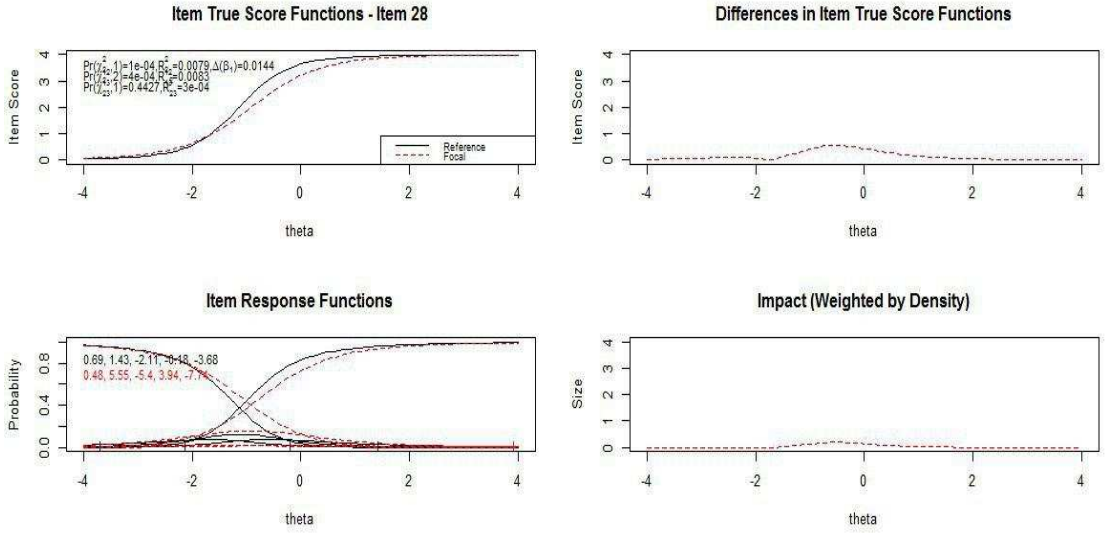

Figure 4. Item 28 plot.

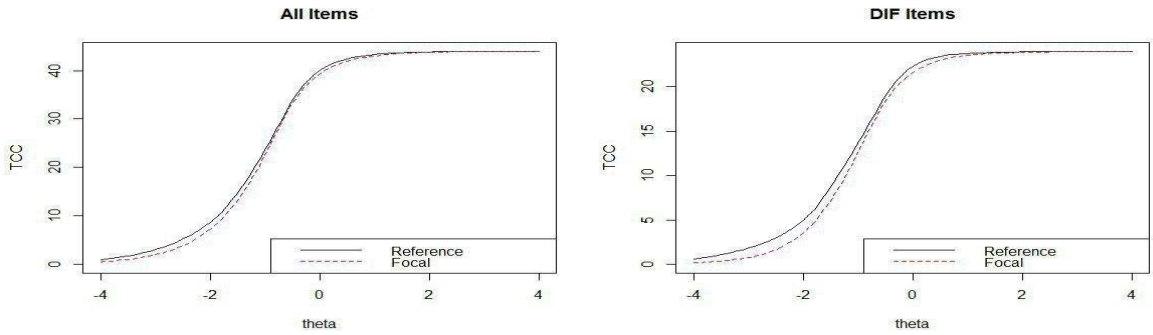

Figure 5. 30 Items plot.

Discussion and Conclusion

Fairness in testing is of paramount importance in educational and psychological testing in order to assess examinees fairly and equitably without bias (Standards for Educational and Psychological Testing, 1999). It can be met in the framework of test development by observing some steps in order to have coherence of claims. However, this issue can rarely be accomplished prior to test administration. According to Kane (2010), "the generic interpretive argument is made during the test development which may carry some sorts of confirmationist bias, because it is an integral part of developing an assessment to support certain interpretations and uses" (p. 181). This would be an indication of content bias known as differential item functioning (DIF) which implies that there are different probabilities among test takers to get an item correct.

In a variety of DIF studies there are three interrelated DIFs including content or language variety, group performance, and standard setting (Kunnan, 2007). In the present research, group performance considers differences in terms of gender performance. In detail, group difference can occur in focal group, and reference group for comparison. To avoid bias, DIF detection is applied in developing tests. Using Lordif software, the results revealed that gender played a significant role in responding to tests. Overall, female participants outperformed their male counterparts although the gap was not that large. At the same time, it was found that all items were not free from DIF. Interestingly enough, DIF continuum carried non–uniform feature that is to say females were thought to be more competent, but male test takers with a

high ability did better on some test items. However, because of the low effect size, this difference might be negligible.

Acknowledgements

The Authors thank *Professor Antony J. Kunnan* for his constructive comments on this piece of work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or non-for-profit sectors.

The Authors

Masoud Geramipour has a Ph.D. in Assessment and Measurement from the Faculty of Psychology and Education at Allameh Tabatabie University, Tehran. Currently, he is an Assistant Professor in the Department of Curriculum Studies, Kharazmi University. He has taught research methodology and psychometrics courses to educational research students since 2010.

Email: mgramipour@yahoo.com

Niloufar Shahmirzadi (corresponding author) is a Ph.D. candidate of Applied Linguistics at Islamic Azad University, Central Tehran Branch. She is a part-time lecturer and has published some articles and books. She has also attended some national and international conferences. She is a member of the Young Researchers and Elite Club. Her areas of interest mainly lie in Language Testing, Assessment, and Educational Measurement.

Email: niloufar_shahmirzadi83@yahoo.com

References

Alderson, J. C., & Urquhart, A. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, 2, 192–204.

Andersen, R. B. (1967). On the comparability of meaningful stimuli in cross-cultural research. *Sociometry*, 30, 124–136.

Angoff, W. H. (1993). *Perspectives on differential item functioning*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Brown, J., D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*, 217–238.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (pp. 221–256). New York: American Council on Education & Praeger series on higher education.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155–163.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016). Lordif: An R package for detecting differential item functioning using iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. (Online PDF Published Manual).

Clauser, E. B., & Mazor, M. K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31–44.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202–14.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and modern Greek. *Language Learning*, *46*, 233–282.

Hale, G. A. (1988). Student major field and text content: Interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing*, *5*, 49–61.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*, 177–182.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*, 89–114.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93–116.

Kunnan, A. J. (2007). Test fairness, test bias, DIF. *Language Assessment Quarterly*, *4*(2), 109–112.

Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, online first*.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357–367.

Ogretmen, T. (2015). DIF analysis across genders for reading comprehension part of English language achievement exam as a foreign language. *Educational Research and Reviews*, *10*(11), 1505–1513.

Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing*, 21, 53–73.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*, 187–210.

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28*(1), 38–49.

Perkins, A. J., Stump, T. E., Monahan, P. O., McHorney, C. A. (2006). Assessment of deferential item functioning for demographic comparisons in the MOS SF-36 health survey. *Quality of Life Research*, *15*(3), 331–348.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371.

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, 9, 12–29.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281-315). Hillsdale NJ: Erlbaum.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Language Testing and Assessment, 4*(1), 97–124.

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

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods. *Educational Measurement, technical report* N. 2.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.