Middle-East Journal of Scientific Research 17 (2): 205-212, 2013
ISSN 1990-9233
© IDOSI Publications, 2013
DOI: 10.5829/idosi.mejsr.2013.17.02.12170

Bias and Differential Item Functioning in Post-Graduate English Proficiency Tests in Iran

Arshya Keyvanfar and Niloofar Dadfarma

Faculty of Foreign Languages,
Islamic Azad University, North-Tehran Branch, Iran

Abstract: Differential item functioning (DIF) has been a recent success in the field of test fairness and or item biasness as it has been broadly used by many researchers worldwide based the present literature as a validity confirmation method. This paper has attempted to briefly introduce and explain the basic issues of DIF in simple terms and provide the reader with a fairly comprehensive review of the recent validation works done by researchers on English Proficiency Test (EPT) in Iran. EPT is a national high-stake test held in Iran for PhD students as a general English proficiency evaluation before they are allowed to be PhD graduates. Therefore, the study has highlighted the strong and weak points of the investigations been ran on the subject as well as that of the test development, validation, administration, scoring and interpretation itself.

Key words: Item Biasness % Differential Item Functioning (DIF) % English Proficiency Test (EPT) % High-Stake Testing % Test Validity % Test Fairness

INTRODUCTION

The history of modern methods of detecting test and item bias as Fletcher (2009) [1] explains started in 1960's with a change of perspective from focusing on test results to standardized assessment in order to create more equity among participants of different backgrounds. Fairness in the field of language testing or what may be referred to as 'ethical' testing, 'unbiased' testing or 'moral issues of testing' is of more significance when it comes to high-stake assessment which is done in a large scale and leaves lifelong effects on test takers' lives [2, 3]. In fact, when one is to select and/or make use of any kind of psychological tests, what is crucial to ascertain is that a test is fair to all the participants of a population and not biased towards some [4]. As mentioned by Geranpayeh and Kunnan (2007) [5], at first relative item-difficulty was the concern of the test bias or Differential Item Functioning (DIF) analysis and it was only after two decades that the focus shifted to the concept of DIF, as it is known nowadays. Holland and Wainer (1993) [6] report that, the history of equity and fair treatment in the Educational Testing Service dates back to the early twentieth century. These concepts root back in the American military and to be exact in its Air Force Human Resource Laboratory. DIF as one of the latest validation tools has become especially important in the Air Force since they need to develop their own tests for different purposes such as hiring, promotion and certification. The term DIF was first used by Holland and Thayer (1988) [7] in their book on validation [8]. As they have stated the early works done on DIF were mainly focused on black and white bias and that it was not until the 1970s that the statistical methods for the detection of potentially biased items came into being. On the other hand, Tatsuoka, Linn, Tatsuoka and Yamamoto (1988) [9], have pointed that studies related to what is contemporarily referred to as DIF dates back to 1951.

In addition to international English proficiency high-stake tests administered in Iran, such as IELTS, TOEFL, GRE and the like; there are different national high-stake tests held in Iran, some of which are of internal use such as Iranian National University Entrance Exam (INUEE) at Bachelor's, Master's and PhD levels in

Corresponding Author: Niloofar Dadfarma, Faculty of Foreign Languages,
Islamic Azad University, North-Tehran Branch, Iran.

different majors and subjects of study. Some general university admission exams are administered for distance learning, like the ones held by Payam-e-Noor and Jame-e- Elmi- Karbordi University, some online courses and some Open University courses at different academic levels. Moreover, there are some tests for internal or external scholarship donations as well as those held for the purpose of employment. Among all, English Proficiency Test, EPT, is a high-stake test, which is held by Islamic Azad University, Science and Research Branch for PhD graduates-AIUEPT. A similar test is also held by Tehran University for public university PhD candidates known as UTEPT, or PhD TOEFL. They currently serve two purposes: one as part of the exit behavior of graduating PhD candidates and one as the entry behavior of those who are to receive scholarships. EPT is also planned to be used for the selection and admission of PhD candidates in the near future. Hence, it seems that EPT as a high-stake test at the postgraduate academic level in Iran is important enough to be studied in terms of its potential to be of equal ease or difficulty for all the test takers [10].

Differential Item Functioning (DIF): Rudas and Zwick (1997) [11] argue that most educational researchers and scholars have considered the absence of differential item functioning as a crucial aspect of test fairness. Robitzsch and Rupp (2009) [12] point out that detection of the potential assessment bias either in item level, Differential Item Functioning, DIF, bundles of groups of items, Differential Bundle Functioning, DBF [13-15], at scale level or the whole test, Differential Test Functioning, DTF [16-18], Differential Passage Functioning, DPF [15], Differential Alternative Functioning, DAF, Differential Distractor Functioning, DDF [19-21], Differential Domain Functioning, DDF [22], Differential Facet Functioning, DFF [23], Differential Testlet Functioning, DTF, has been the concern of great many researches and studies. In the present study, the researchers intend to focus on individual item level and specifically on Differential Item Functioning (DIF) known as a classical method of identifying flagged biased items.

Differential item functioning (DIF) being categorized as a part of test score validity, is a quantitative characteristic used for the evaluation of potential item bias [24]. DIF, as Robitzsch and Rupp (2009) [12] suggest, "refers to the fact that statistically significant differences in item operating characteristics, such as item difficulty, exist across subgroups of examinees that have been matched on some valid measure of the construct that the test is measuring" (p. 19).

As it is believed by the majority of researchers such as [13, 25-27, 12], the presence of DIF could indicate that an item is biased towards a group of applicants who are of equal ability save for one. They believe that this difference might be ascribed to some kind of ethical, ethnical, racial, cultural, linguistic, socioeconomic, or academic background, or there may be differences in age, gender, disability and nationality of the competing groups. This difference in performance is revealed when the test takers are compared against a matching criterion which serves as the grouping factor. Karimi (2011) [28] mentions that this matching process is also called conditioning process. He further adds that this criterion could be either internal to the test (for example, the total score of the test, raw scores, the rest score of the test without the item under investigation or the latent score) or external (for instance, a performance measure such as a school grade, total score of another test, etc.). It is worth mentioning that, as mentioned by Karimi and Salmani Nodoushan (2011) [29], the latter should be done with caution for one should be concerned with the unbiasedness of the supplementary test used for the external matching and that it is measuring the same construct as the main test.

The occurrence of DIF is not necessarily a sign of biasness unless the concept being measured is a secondary dimension not meant to be measured and hence is inevitably considered to be irrelevant and nuisance. This way DIF is called to be 'adverse' and is assumed to be a real threat to the validity of the test as a whole, the proceeding results and/or decisions made based on the findings [29]. The reason why this difference could pose a threat to the validity is that it causes systematic errors in the following inferences. Therefore, from one point of view, this could be seen as a flaw; since the detection of DIF in itself is not enough to prove the biasness of test and there is an obligatory need for further analysis by experts. This item analysis could also be considered as another source of weakness for it is too subjective and needs to be reviewed by a number of test experts in order to increase the validity level of this action [29]. On the contrary, if the secondary construct is intended to be measured or relevant, it is called 'auxiliary' and it is displaying a 'benign' DIF or impact, which is desirable [13, 27]. There are different methods of detecting DIF some of which are Logistic Regression (LR), Mantel-Haenszel (MH) Odd Ratio Chi-Square, Item

Response Theory (IRT), Structural Equation Modeling (SEM), Logistic Discriminant Function Analysis (LDFA), Multiple Indicators Multiple Causes (MIMC), Mean and Covariance Structure (MACS), Correct Response Rates (p-value), Transformed Item Difficulty Index (TID) or delta plot a.k.a. Transformed Item Difficulty (TID), Standardization procedure (STND) and Simultaneous Item Bias Test (SIBTEST). Alavi and Karami (2010) [30] have mentioned that some of the techniques such as the delta method are no longer employed due to their computational limitations. On the other hand, IRT based methods though stylish are statistically and computationally very complicated in comparison to other non-IRT based methods. They have cited that in non-IRT based techniques, test takers are matched based on an observed variable; however, in IRT based approaches, they are matched based on the latent trait. Depending on the method of detecting DIF, there could be two or more groups of participants, but usually there are two groups under the titles of a 'focal group' and a 'referent group'[31]. De Ayala (2009) [32] as well as Linacre and Wright (1989) [33] elaborate that the focal or the 'minority' group is the one under investigation to see if it shows any differences in performance. As Karimi and Nodoushan (2011) [29] point out, the focal group is the "potentially disadvantaged group" whereas the referent group is the "potentially advantaged group by the test"; meanwhile, they pinpoint that this grouping could be arbitrarily done at times. The referent group is called the 'majority' or 'comparison' group by De Ayala (2009) [32] and 'target' group by Geranpayeh and Kunnan (2007) [5] as it is this group against which the standard performance of the items of interest could be compared.

Uniform and non-uniform DIFs are two forms of DIF. In uniform or unidirectional DIF, as De Ayala (2009) states, throughout the continuum, "one group performs better" (p.14). According to Pae and Park (2006) [34], uniform DIF could be defined in terms of items showing probability differences in item difficulty. In non-uniform or crossing DIF, De Ayala (2009) [32] explains, for a particular portion of the continuum "the referent group performs better than the focal group; along a different portion of the continuum the focal group outperforms the referent group" (p. 343). On the other hand, Pae and Park (2006) [34] believe that nonuniform DIF could be somehow similar to item discrimination.

EPT: IAUEPT (Islamic Azad University English Proficiency Test) or EPT as a national test is being administered since late 1986 by Islamic Azad University, Science and Research Branch four times a year. UTEPT (University of Tehran English Proficiency Test) or Doctorate TOEFL had been established few years ahead of IAUEPT and is also held four times a year by Tehran University [35]. They are both designed to measure the English general proficiency level of Iranian PhD candidates. In fact, they are used as a proficiency test along with other requirements to enter and exit the PhD program during admission and graduation respectively.

IAUEPT consists of 100 multiple-choice items in three subtests: 20 vocabulary items, 40 grammar items (20 structure and 20 correction written expressions items) and 40 reading comprehension items (four passages each containing 10 reading comprehension questions). The test has to be taken in 120 minutes time. UTEPT, on the other hand, as Amiran (2012) [35] explains, is comprised of 35 grammar items (15 structure, 10 written expressions and 5 grammar in context items), 35 vocabulary items (30 synonyms and 5 vocabulary in context items), 30 reading comprehension items (6 passages with 4-8 questions summing up to 26 reading comprehension questions and 4 restatement items). Unlike IAUEPT, this test allocates only 100 minutes time to its 100 items.

Test Development: The Assessment Center of IAU (Markaze Azmoon) is the organization in charge of the construction, validation and administration of EPT which seems to be a tailored version of the paper-based TOEFL test. We were informed during some personal communications that three different sets of items are constructed by the team of item developers and then one set is randomly selected by a different team for security purposes as the final version. Although high stake tests undergo prototyping (that is,the process of testing the function of newly developed assessment tools) before being used in test takers' assessment projects [36], to the best of our knowledge, the Assessment Center does not have the luxury of prototyping EPT leaving its validity open to question.

Administration and Scoring: The exam session starts with the demographic questions such as age, gender, academic field of study and university. The test is administered along with a paper answer sheet.

As the tests are multiple-choice type with four options, they can be simply scored by machines. That is to say, the answers filled in the answer sheets are transferred into computers and then scored in the form computerized data. The cutting score of 50 out of 100 is the passing mark. The scoring system as Karami (2012)

[37] has mentioned is dichotomous and the items which are not answered are considered as being wrong. It is worth mentioning that the test unlike other national tests in Iran does not assign negative marks to missing or wrong answers.

As Pae (2004), points out, there are few studies on DIF outside the States and even fewer on fields of study. In particular, few studies have examined the methods of flagging DIF items for tests administered to Iranian university students of different academic backgrounds.

DIF Studies in Iran: A basic problem often associated with decision-making particularly in the field of high-stake testing is the issue of being ‘unbiased’. This becomes particularly an issue of concern when test-takers come from different backgrounds and hence introduce diverse characteristics to the evaluation setting [38]. Therefore, one could see the need to study different tests, especially those that are of grave importance to the test takers’ professional and or academic lives. The following is a brief account of the studies conducted to investigate test bias or DIF at item level in the English Proficiency Test PhD candidates in Iran have to sit for some time along their course of study and obtain a minimum of English general proficiency before they graduate.

Ali Rezaee and Shabani (2010) [8] have examined gender bias in November 2006 version of the University of Tehran English Proficiency Test (UTEPT) with a sample of 6555 test-takers of different master majors out of which 69.5% were male (referent group) and 30.5% were female (focal group) master holders.. DIF was measured using multistep Logistic Regression and to ensure the significance of the DIF, a two-degree freedom Chi-square test and to check for the effect size an R-square test were employed. The results revealed that 39 out of 100 items showed gender DIF, most of which were in favor of the male group especially in the Reading Comprehension part. Based on the effect size, however, they were reported as negligible and hence the test was not considered to show gender differences.

In another study, Alavi and Karami (2010) [30] make their own interpretations of DIF items. They introduce three main problems. The first problem they raise is the question of if interpretations of the results gained using DIF are allowed to be applied in the sense of fairness detection. The answer to this question seems to be no. They argue that the method is only capable of detecting the difference in the performance of the groups and if this were proven, it would not by any means function as the evidence for the presence of bias. They also address the issue of choosing a valid matching criterion. They state that since this criterion is usually the total or trait test score, then the criterion could not have been free of that trait if the test is identified as biased. They also refer to another wise possibility explaining that if the majority of items are showing DIF, then ones which are neutral could be viewed as functioning differentially toward both groups. The last issue they show their concern about is the subjective judgment of test experts which is the most vital stage of DIF and is called ad hoc interpretations. This study tries to examine the role of test expert in the interpretation of DIF results. Here 5336 students (68.5 % Humanities and 31.5% Science and Technology group) took the PhD acceptance proficiency test, University of Tehran English Proficiency Test, UTEPT. Only the 25 items of the vocabulary section were analyzed in this study. It was found that 14 items displayed DIF, seven in favor of Humanities and the other seven in favor of the Science and Technology group. A questionnaire was also utilized to gauge the DIF items. It comprised of two sets of six items, each in favor of one group. At last it was revealed that in each set there were two items in favor of the group of interest, two neutral items and two against the group of interest. The IRT Rasch model was used to detect DIF items. Their expert panel comprised of two TEFL PhD holders, four PhD candidates and four MAs. The analyss of test experts’ opinions showed that there was not agreement of any kind, inter or intra, among them. That is some experts focused on stems to locate the source of DIF while others focused on the options or that they shifted their focus from stem to options or the other way round at times. However, the interesting point was that there was no reason behind this and no firm logical theoretical basis could support their moves. Hence, as the authors expressed, it was more likely that the testing specialists were constantly shifting from one strategy to another just to justify the possible existence of DIF, if any. This can be the same as ad hoc interpretation which is empirically thought to add nothing to the basis and at best succeeds to support the underlying theory [30]. There were other contradictions as well; to name some one could mention that at times experts tried to come up with explanations for DIF for items that did not display DIF; some items were easily detected just by being read and the DIF analysis was presumably of no point for them. In some case, there was obvious evidence of doubt as the expert could not decide if the item was in favor of this group or that one and attempted to guess the predictability. One last point mentioned by Alavi and

Karami (2010) [30] for basing a firm logical and theoretical foundation for experts' analysis is the idea of exercising the Think Aloud Protocol (TAP) to experts when judging which has been applied by Ercikan, Arim, Law, Domene and Lacroix (2010).

Karami and Shabani (2011) [10] have compared the results of Mantel-Haenszel and the Rasch model DIF detection techniques implementing the same data used by Karami (2011) [28]. The applicants consisted of two main academic backgrounds, Humanities and Science and Technology groups. Only one item was flagged as showing DIF by the MH method, which was negligible taking the whole test into account. Therefore, it could be concluded that the result of either of the methods are fairly consistent and comparable. Moreover, these different methods could be viewed as complementary approaches rather than rivals [10].

Karami (2011) [28] has scrutinized the effect of gender on 1562 PhD applicants as test-takers (63.4% male and 36.6% female) of the University of Tehran English Language Proficiency Test (UTEPT). In this version of this test, 19 items were flagged as DIF among which only three items were detected as displaying partial DIF. In spite of these, none of the spotted DIF items had a biased source. However, there remains one item, which was mistakenly repeated twice and had probably caused the loss of two points for the ones who were ignorant of them and that is why the author believes that this questions the fairness of such a high-stake test.

Alavi, Ali Rezaee and Amirian (2011) [39] have implemented generalized Mantel-Haenszel and Logistic Regression to spot DIF items. A sample of 400 master holders in the academic fields of Humanities (social sciences, law, political sciences, management, Persian literature and foreign languages), as the referent group and Science and Engineering (chemistry, physics, mathematics, biology, agricultural, mechanical, electrical and civil engineering) as the focal group were studied. They all had taken part in the University of Tehran English Proficiency Test (UTEPT) for admission to PhD. In the descriptive statistics section, it was discovered that although the Science and Engineering group's performance was slightly better than the other groups', this difference was not meaningfully significant. The application of GMH resulted in the identification of 12 DIF out of 100 items, but Logistic Regression analysis flagged 14 items. Eventually due to the negligible effect size none of them was considered biased towards any of the groups.

Karami (2012) [37] has sought the possible effect of persons, items, subtests and academic background on the test's fairness and the dependability of scores from University of Tehran English Proficiency Test (UTEPT) through the use of Generalizability Theory (G-theory) and Classical Test Theory (CTT) reliability analysis. All 5795 PhD graduating candidates who sat for the test in 2004 were form four academic majors of Agriculture (13.4%), Humanities (58.1%), Science (13.2%) and Technology (15.3%). Using two-level sampling, out of this population a sample of 3068 test-takers was randomly selected, so that an equal number of 767 participants was allocated to each group. To have the same number of items in each subtest, 25 items were randomly selected from the grammar and reading comprehension sections. The findings indicated that through both aforementioned statistical methods, the test showed a rather high reliability level and that the test was not biased against any academic background groups.

Amiran (2012) [35] has investigated academic background DIF items across UTEPT test using Mantel-Haenszel (MH) method. The participants were 1550 PhD students who took the test in 2010. They came from two main academic backgrounds: 809 Humanities (referent group) and 741 Science and Engineering test-takers (focal group). A questioner as the scale of judgment for biased DIF items was handed to the panel of experts. It was revealed that 13 items displayed DIF (3 grammar, 7 vocabulary, 3 reading comprehension items) and after the calculation of effect size by ETS (Educational Testing Service), only 4 items (3 vocabulary items in favour of Humanities and 1 grammar item in favor of Science and Engineering group) showed moderate DIF. The content analysis done by two experts, after being briefed on the scale, confirmed the bias in the items. On the whole it was concluded that only 4 out of 100 were biased indicating the acceptability of the test and its fairness to both of the academic background groups.

Keyvanfar and Dadfarma (2012) [40] have recently made an attempt to detect differentially functioning items utilizing Mantel-Haenszel and Logistic Regression procedures in IAUEPT across a sample of 1033 test-takers (61.9% males, focal group, 38.1% females, referent group; 53.3% Social Sciences, focal group and 46.7% Technical Sciences and Engineering, referent group). The results indicated there were no items flagged as displaying gender DIF by either of the methods. After purification processes and computation of effect size, MH revealed 10 moderate fields of study uniform DIF items; eight items favored the referent group while only two items favored the focal group. The number of items flagged by MH and LR procedures was different. This study was unable to decide if the type of items flagged could have been different, since no items were flagged by LR. Further investigations also revealed that item format module and content, did not make it any easier for either of the groups to answer items correctly therefore presence of bias was not confirmed.

Sample biased DIF items: The followings are the Biased items detected in Amiran's (2012, p. 7-10) [35] study.

Item 33: The emergence of endocrinology as a separate discipline can be ……32…….. This …....33…… is secreted from cells in the intestinal walls when food …34…… the stomach.

A. substance**T**B. sprayC. solutionD. subject

Item 40: There is evident conflict between Henry's social philosophy and the actions of his character.

A. generalB. obvious**T**C. importantD. appropriate

Item 50: As *paradoxical* as it may seem, the infinity of even numbers is as big as that of all numbers.

A. mathematicalB. contradictory**T**C. emotionalD. rudimentary

Item 52: Many people in the past could neither read nor write. They were ……………..

A. illiterate**T**B. traditionalC. emotionalD. cultural

While item 33 is in favor of Science and Engineering, the other three items have displayed bias against this group [41-43]. After the content analysis done in this study, Amiran (2012) [35] suggests that since the nature of the text for the first item is scientific, it is more probable for Science and Engineering group to have seen this word in their textbooks while the other items are more probable to be seen in humanities contexts.

CONCLUSION

As the few examples mentioned here and other works in the literature suggest, DIF is among one of the common methods which is currently used to detect and remove the bias from the tests. There are quite a wide range of statistical methods to be used for flagging DIF items some of which were listed previously here; however, it is the responsibility of the researcher to choose the best matching method for his studies among all the various techniques and methods due to the nature of test, research, practicality, etc as they are suggested for each method. It is noteworthy that the last faze of bias detection utilizing DIF which is the experts' analysis needs to be done with optimum care since this is the critical step which is going to decide if a flagged DIF item is actually biased and consequently is going to be removed or not. This paper has gathered a rather comprehensive collection of the studies ran on EPT as a prominent high stake test held for Iranian PhD students, so that it can provide the test makers, policy makers, test experts, researchers, etc a holistic view on the validation and standardization processes and or studies done recently. This study aimed to highlight the importance of the test itself and also signal the paramount necessity of execution of optimum care towards all the steps taken throughout the whole process of testing namely test development, piloting, editions and revisions, standardization, validation, administration, scoring and interpretations. This review has also attempted to be to the benefit of the researches in that they could find the niche as it is untouched or of essential enough to be studied as well as the idea of a follow up study which seems to be pretty much missed with this regard.

REFERENCES

1. Fletcher, J., 2009. Detecting Differential Item Functioning (DIF) in the diabetes risk perception survey (RPS-DM). ETD Collection for Fordham University, pp: 1-143.
2. Salehi, M. and A. Tayebi, 2012. Differential item functioning: implications for test validation. Language Teaching and Research, 3(1): 84-92.
3. Barati, H., S. Ketabi and A. Ahmadi, 2006. Differential Item Functioning in high- stake tests: The effect of field of study. IJAL., 9(2): 27- 49.
4. 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 ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
5. Geranpayeh, A. and A.J. Kunnan, 2007. Differential Item Functioning in terms of age in the certificate in advanced English examination. Language Assessment Quarterly, 4(2): 190-222.

6. Holland, P.W. and H. Wainer, 1993. Differential Item Functioning. NJ: Lawrence Erlbaum Inc., Publication.
7. Holland, W.P. and D.T. Thayer, 1988. Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity, pp: 129-145. Hillsdale, NJ: LEA.
8. Ali Rezaee, A. and E. Shabani, 2010. Gender Differential Item Functioning Analysis of the University of Tehran English Proficiency Test. Pazhuhesh-e Zabanha-ye Khareji, 56: 89-108.
9. Tatsuoka, K.K., Linn, L. Robert, M.M. Tatsuoka and K. Yamamoto, 1988. Differential item functioning resulting from the use of different solution strategies. Journal of Educational Measurement, 25: 301-19.
10. Karami, H. and E.A. Shabani, 2011. On the comparability of two DIF detection techniques: Mantel-Haenszel and the Rasch model. The 9th International TELLSI conference, October, pp: 20-22, Ilam, Iran.
11. Rudas, T. and R. Zwick, 1997. Estimating the importance of differential item functioning. Journal of Educational and Behavioral Statistics, 22: 31-45.
12. Robitzsch, A. and A.A. Rupp, 2009. Impact of missing data on the detection of differential item functioning: the case of Mantel-Haenszel and Logistic Regression Analysis. Educational and Psychological Measurement, 69(1): 18-34.
13. Abbott, M.L., 2007. A confirmatory approach to Differential Item Functioning on an ESL reading assessment. Language Testing, 24(7): 7-36.
14. Liu, O.L., M. Schedl, J. Malloy and N. Kong, 2009. Does content knowledge affect TOEFL iBT reading performance? A confirmatory approach to Differential Item Functioning. TOEFL iBT Research Report-09. ETS.
15. Hill, Y.Z. and O.L. Liu, 2012. Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT reading performance? TOEFL iBT Research Report-18. ETS.
16. Runnel, J., 2013. Measuring Differential Item and Test Functioning across academic disciplines. Language Testing in Asia, 3: 9.
17. Gong, J., 2012. Detection of Differential Test Functioning (DTF) and Differential Item Functioning (DIF) in MCCQE part II using logistic models.
18. Bolt, D. and W. Stout, 1996. Differential Item Functioning: Its multidimensional model and resulting SIBTEST detection procedure. Behaviormetrika., 23(1): 67-95.
19. Middleton, K. and C.C. Laitusis, 2007. Examining Test Items for Differential Distractor Functioning among students with disabilities. ETS Research Report. ETS.
20. Abedi, J., S. Leon and C.J. Kao, 2008. Examining Differential Distractor Functioning in reading assessment for students with disabilities. CRESST Report 743. CSE.
21. Kato, K., R. Moen and M. Thurlow, 2007. Examining DIF, DDF and omit rate by discrete disability categories. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.
22. Zumbo, B.D. and M.N. Gelin, 2005. A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/ community moderated (or mediated) test and item bias. Journal of educational research and policy studies, 5(1): 1-24.
23. Yi, D., B.D. Wright and W.L. Brown, 1996. Differential Facet Functioning detection in direct writing assessment. ERIC Reports, pp: 1-24.
24. Wood, W.S., 2011. Differential item functioning procedures for polytomous items when examinee sample sizes are small. Doctoral Dissertation. University of Iowa. retrieved at http://ir.uiowa.edu/etd/1110
25. Birjandi, P. and M. Amini, 2007. Differential Item Functioning (Test Bias) Analysis Paradigm across Manifest and latent examinee groups (on the construct validity of IELTS). Human Sciences, 55: 1-20.
26. Kim, M., 2001. Detecting DIF across the different language groups in a speaking test. Language Testing, 18(1): 89-114.
27. Pae, T.L., 2004. DIF for examinees with different academic backgrounds. Language Testing,21(1):53-73
28. Karami, H., 2011. Detecting gender bias in a language proficiency test. International Journal of Language Studies (IJLS), 5(2): 27-38.
29. Karami, H. and M.A. Salmani Nodoushan, 2011. Differential Item Functioning (DIF): current problems and future directions. International Journal of Language Studies (IJLS), 5(3): 133-142.
30. Alavi, S.M. and H. Karami, 2010. Differential Item Functioning and ad hoc interpretations. TELL, 4(1): 1-18.
31. Barr, M.A. and N.S. Raju, 2003. IRT-based assessments of rater effects in multiple-source feedback instruments. Organizational Research Methods, 6(1): 15-43.
32. De Ayala, R.J., 2009. The theory and practice of item response theory. NY: Guilford Press.
33. Linacre, J.M. and B.D. Wright, 1989. Mantel-Haenszel DIF and PROX are equivalent. Retrieved at www.rasch.org/rmt/rmt32a.htm

34. Pae, T.L. and G.P. Park, 2006. Examining the relationship between differential item functioning and differential test functioning. Language Testing, 23(4): 475-496.
35. Amiran, S.M.R., 2012. Investigating academic discipline bias in UTEPT using MH method. The 1st Conference on Language Learning & Teaching (An Interdisciplinary Approach) October, 2012, Ferdowsi University of Mashhad, Mashhad, Iran.
36. Azeem, M. and B.M. Gondal, 2011. Prototype framework: prototypes, prototyping and piloting in terms of quality insurance. Academic Research International, 1(2): 301-308.
37. Karami, H., 2012. The relative impact of persons, items, subtests and academic background on performance on a language proficiency test. Psychological Test and Assessment Modeling, 54(3): 211-226.
38. Lochenr, K. and A. Preuss, 2012. Valuing diversity through fair testing. Cut-e Group. White Paper., 1(5): 1-15.
39. Alavi, S.M.A. Ali Rezaee and S.M.R. Amirian, 2011. Academic discipline DIF in an English language proficiency test. Journal of English Language Teaching and Learning, 5(7): 39-66.
40. Keyvanfar, A. and N. Dadfarma, 2012. Differential Item Functioning (DIF) of Field of Study and Gender in English Proficiency Test (EPT) of Iranian PhD Candidates at IAU Tehran Research and Science Branch. European Journal of Scientific Research, 18(1): 132-143.
41. Hossein Berenjeian Tabrizi, Ali Abbasi and Hajar Jahadian Sarvestani, 2013. Comparing the Static and Dynamic Balances and Their Relationship with the Anthropometrical Characteristics in the Athletes of Selected Sports, Middle-East Journal of Scientific Research, 15(2): 216-221.
42. Anatoliy Viktorovich Molodchik, 2013. Leadership Development: A Case of a Russian Business School, Middle-East Journal of Scientific Research, 15(2): 222-228.
43. Meruert Kylyshbaevna Bissenova and Ermek Talantuly Nurmaganbet, The Notion of Guilt and Problems of Legislative Regulations of its Forms: The Notion of Guilt in the Criminal Law of Kazakstan, Middle-East Journal of Scientific Research, 15(2): 229-236.