Content list available at www.urmia.ac.ir/ijltr

Iranian Journal of Language Teaching Research

Urmia University

Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran

Alireza Ahmadi [a, *], Ali Darabi Bazvand [a]

[a] *Shiraz University, Iran*

ABSTRACT

Differential Item Functioning (DIF) exists when examinees of equal ability from different groups have different probabilities of successful performance in a certain item. This study examined gender differential item functioning across the PhD Entrance Exam of TEFL (PEET) in Iran, using both logistic regression (LR) and one-parameter item response theory (1-p IRT) models. The PEET is a national test consisting of a centralized written examination designed to provide information on the eligibility of PhD applicants of TEFL to enter PhD programs. The 2013 administration of this test provided score data for a sample of 999 Iranian PhD applicants consisting of 397 males and 602 females. First, the data were subjected to DIF analysis through logistic regression (LR) model. Then, to triangulate the findings, a 1-p IRT procedure was applied. The results indicated (1) more items flagged for DIF by LR than by 1-p IRT (2) DIF cancellation (the number of DIF items were equal for both males and females), as revealed through LR, (3) equal number of uniform and non-uniform DIF, as tracked via LR, and (4) female superiority in the test performance, as revealed via IRT analysis. Overall, the findings of the study indicated that PEET suffers from DIF. As such, test developers and policymakers (like NOET & MSRT) are recommended to take these findings into serious consideration and exercise care in fair test practice by dedicating effort to more unbiased test development and decision making.

Keywords: differential item functioning; logistic regression; one-parameter item response theory

ARTICLE HISTORY

Received: 18 Nov. 2014 **Revised version received:** 22 Dec. 2015

Accepted: 23 Dec. 2015 **Available online:** 1 Jan. 2016

* Corresponding author: Department of Foreign Languages & Linguistics, Faculty of Literature & Humanities, Shiraz University, Shiraz, Iran
Email address: arahmadi@shirazu.ac.ir

Introduction

Differential item functioning (DIF) has generated great interest in language testing applications (see Holland & Wainer, 1993; Penfield & Camilli, 2007). When equally knowledgeable subgroups of examinees (e.g., gender groups) exhibit differential probabilities of a correct response for an item, differential item functioning (DIF) is the result(e.g. Angoff, 1993; Camilli & Shepard, 1994; Pae, 2012; Zumbo, 2007). Researchers believe that through the use of DIF detection methodologies, factors contributing to group differential performance could be revealed, items flagged for DIF could be discarded, and finally fairer decisions could be made (Pae, 2004a; Rezaee & Shabani, 2010).

For this reason, literature is chock-full of different DIF detection studies conducted within various testing programs and across different sub-groups of examinees. These studies have focused on factors such as language background (e.g., Harding, 2011; Kim & Jang, 2009), ethnicity (e.g., Stoneberg, 2004), age (e.g., Geranpayeh & Kunnan, 2007), linguistic backgrounds (e.g., Ryan & Bachman, 1992), academic background (e.g., Pae, 2004a), disability status (e.g., Maller, 1997), text familiarity (e.g., Ahmadi & Jalili, 2014; Pae, 2004b), field of study (Barati, Ketabi, & Ahmadi, 2006; Näsström, 2004), and finally gender which, due to its significant role in high stakes decision making, has attracted a clear majority of some well-documented studies in the field (e.g., Amirian, Alavi & Fidalgo, 2014; Li & Suen, 2013; Pae, 2012; Song, Cheng, & Klinger, 2015).

Across such gender-related studies, two lines of research emerge in the pertinent literature. The first includes studies examining gender DIF within tests of language proficiency in general (e g., Amirian, Alavi & Fidalgo, 2014; Song, Cheng, & Klinger,2015), and the second studies investigating gender DIF across subject-matter tests such as mathematics (e.g., Mendes-Barnet & Ercikan, 2006) and science (Zenisky, Hambleton & Robin, 2003).

Differential performance on tests of subject matter tests (used as entrance tools) that favor one group of examinees over another enhances our understanding of DIF and accordingly brings the task of decision making to a sort of evenhandedness; therefore, questions arise as to whether item characteristics that favor one group of examinees (male or female) within tests of language proficiency, mathematics and science show themselves within high stakes tests of subject matter such as TEFL, as well. Few studies (with no study on TEFL) have been carried out to investigate gender DIF through subject-matter tests (e.g., Gierl, Bisanz, Bisanz, & Boughton, 2003; Mendes-Barnett & Ericikan, 2006; Pae, 2012). Accordingly, more empirical studies regarding the possible gender DIF in high-stakes content tests are warranted. As such, this study tried to track the presence of gender DIF in the specialized part of PhD Entrance Exam of TEFL in Iran (PEET), by using both one-parameter IRT and Logistic regression (LR) procedures.

Literature review

Previous research on gender DIF

A close look at language testing history shows that considerable attention has been paid to conducting research on DIF in general. Specifically, a number of gender DIF studies have been carried out in testing literature, the sketch of some of which is reviewed here. For example, Ryan and Bachman (1992) found gender differences across Test of English as a Foreign Language (TOEFL) and First Certificate of English (FCE) using Mantel-Haenszel (MH) procedure. With regard to TOEFL, four of the items favored males and two items were biased toward females. As regards the FCE, one item favored males and the other one in favor of females. In the same line,

Amirian, Alavi and Fidalgo (2014) detected gender DIF in a language proficiency test in Iran known as University of Tehran English Proficiency Test (UTEPT) using Mantel-Haenszel and Logistic Regression (LR) methods. Results indicated that 28% of the items displayed DIF, suggesting that humanities related topics were more in favor of females, while science oriented texts were biased for males.

Gender DIF studies across reading comprehension tests have also attracted the attention of researchers. Using MH procedures, Pae (2004b) detected gender DIF across the English subtest of Korean College Scholastic Ability Test (KCSAT) and found that logical inference items were more likely to favor males, while items dealing with impressions, mood and tone of a given passage tended to favor females. Similarly, Pae (2012) systematically examined the same sub-test but on a long term basis and across three regular forms (1999, 2003, 2007), applying MH procedures and IRT-LR methods. It was reported that item type is a more reliable predictor of gender DIF than item content, thus being consistent with his previous (2004b) study. Ahmadi and Jalili (2014) also applied two DIF detection methods of LR and IRT across an Iranian reading comprehension test. Consistent with Pae (2004b, 2012), this study revealed that 17% of the items displayed DIF, suggesting that item types such as reference and vocabulary were better predictors of gender DIF (mostly favoring females) than test content.

Investigating gender DIF within listening tests has also been of concern for researchers. Park (2008) used MHDIF detection across the English listening part of 2003 Korean College scholastic Ability Test (KCSAT). It was revealed that 13 out of 17 items were flagged for gender DIF but with somehow equal proportion for males and females. It was also shown that item content was a better predictor of gender DIF. This finding was inconsistent with the findings of the previous studies as item type rather than item content was reported to be a better predictor of gender DIF (Ahmadi & Jalili, 2014; Pae, 2004b, 2012). Another similar DIF study conducted within test application context of listening skill was that of Aryadoust, Goh, and Lee (2011) in which they applied a t-test uniform and non-uniform DIF analysis on the listening part of Michigan English Language Assessment Battery (MELAB) across gender groups. Uniform DIF analysis indicated two DIF items favoring different gender groups, while non-uniform DIF analysis indicated several DIF items mostly favoring low-ability male test takers.

Using the IRT 1- Parameter Logistic Model (OPLM), Takala and Kaftandjieva (2000) studied gender DIF in the vocabulary subtest of the Finnish Foreign Language Certificate Examination (FFCE), suggesting that no bias in terms of gender effect was displayed in this study.

With regard to verbal ability, findings are in contradiction. For example, Cole (1997) found that girls have a better performance on items measuring verbal ability; Nevertheless, Hyde and Lynn (1988) did not find any differences between males and females in this regard.

As regards the second stream of gender research which is pertinent to subject matter studies, some investigations have been made. In a study of SAT mathematics test, Harris and Carlton (1993) found that abstract algebra items and items requiring low cognitive processing favored females whereas on geometry, measurement, number, computation, data analysis, and proportional reasoning items DIF favored males. Later on, however, Mendes-Barnett and Ercikan (2006) came to the conclusion that boys performed better on items requiring problem solving, high cognitive complexity, visual reasoning, and application of mathematics principles to word problems. Other researchers have identified no systematic gender DIF for mathematics items across different testing application contexts such as California Achievement Tests (Haeok, 1990), and Iowa Test of Basic Skills mathematics problem solving and mathematics concepts items (Plake, 1980).

As regards gender DIF in science, many studies have been carried out in the field. For example, some have examined item format effect (Bolger & Kellaghan, 1990; Hamilton, 1999; Zenisky, Hambleton, & Robin, 2003), suggesting that multiple-choice items seem to benefit males, while open-ended items are more biased for females. Others have studied the effect of item contents (Becker, 1989; Burkam, Lee & Smerdon, 1997; Jovanovic, Solano-Flores, & Shavelson, 1994; Young & Fraser, 1994), concluding that males seem to outperform females on physical, earth, and space science items. Consistently, items requiring spatial reasoning or visual content favored males (Halpern, 1992).

So far, few studies have empirically examined gender-related differences across language proficiency tests in an Iranian context (e g., Ahmadi & Jalili, 2014; Amirian, Alavi & Fidalgo, 2014; Rezaee & Shabani, 2010). When it comes to subject-matter tests, no DIF study has been carried out in the field of TEFL in the Iranian or non-Iranian context. Therefore, a systematic approach to the identification of the potential gender DIF on high stakes subject-matter tests is needed. As such, the present study aimed at bridging this gap. Two research questions guided the study in this regard:

RQ1. Does group membership (gender) have any effect on the performance of PhD applicants across PhD entrance exam of TEFL, as investigated by LR and 1-p IRT?

RQ 2. To what extent does the gender DIF results from LR and 1-p IRT methods correspond?

Methods

Participants

Each year, over 150000 PhD applicants compete for the PhD entrance examinations in Iran. This study analyzed data from all PhD applicants (n=999) who took part in PEET in January 2013, regardless of whether they were subsequently admitted to PhD programs. Female participants (n=602) were specified as reference group, while male (397) participants were considered as focal group. The participants' test performance data was provided by the National Organization for Educational Testing (NOET) at the request of Shiraz University. With regard to DIF studies which apply logistic regression method, a sample size of 200 per group is generally suggested to add power to results and to avoid inflated Type I error (Güler & Penfield, 2009; Mazor, Clauser, & Hambleton, 1992; Paek & Wilson, 2011; Zumbo, 1999). For IRT analysis, depending on the model used, a range of 100 to 1000 is suggested. As such, the group sample size selected for this study was considered as thoroughly adequate.

Instrument

PhD entrance examinations in Iran play a great role in the admission decisions of Post graduate studies. These high stakes examinations consist of a series of centralized written examinations designed to provide information on the eligibility of PhD applicants (with different academic majors) to enter PhD programs. Since 2011, these instruments superseded the traditional university-based examination sets in Iran. They are Multiple-Choice tests administered by the NOET—a central testing organization for preparing, organizing, and scoring the university entrance examinations (UEEs), also known for administering Standard Tests of English such as TOEFL, IELTS and GRE (Kiany, Shayestefar, Ghafar Samar, & Akbari, 2013). Each designed exam consists of three blocks: general competence section, academic talent test and domain-specific section, all appearing in MC format with four-item options.

For this study, the field-specific section of Teaching English as a Foreign Language (TEFL) exam administered in January, 2013 was considered. This field relevant exam which is aimed at measuring the candidates' expertise in the field of TEFL is supposedly related to the courses students have passed in the MA or even BA program. In fact, it assesses the students' domain- related knowledge in areas which are assumed to be the prerequisite for entering the PhD programs since the PhD program is built on such areas of knowledge. As such, the knowledge test of PEET consists of 100 items including questions on linguistics (15 items), foreign/second language teaching methods (15 items), research methods (15 items), language assessment (15 items), theories and issues of language learning and teaching (30 items), and finally sociolinguistics and discourse analysis (10 items). Based on a criterion (cut-off score) determined and decided by NOET, more than three times as many applicants as universities can accept are introduced to the respective universities to be interviewed. The interview questions are related to the participants' research backgrounds, academic records, and expertise (technical knowledge). The final admission will be based on the aggregate scores from the PhD entrance exam in written form and the oral interview.

Analyses

It has been emphasized that employment of more than one method of DIF analysis in DIF investigations may contribute to more dependable results (Aryadoust, Goh, & Kim, 2011; Camilli, 2006; Fidalgo, Alavi & Amirian, 2014; Pae, 2012; Uiterwijk & Vallen, 2005). As such, two methods of DIF detection were applied in this study: First, 1-p item response theory (IRT) model was applied, then as a classical method, logistic regression method was used. The results were then compared to determine the degree of correspondence between the two methods.

DIF analysis via logistic regression. Logistic regression has been widely considered as one of the best statistical methods for investigating DIF (Zumbo, 1999); nonetheless, applying the right strategies of this approach to DIF investigations has been virtually misguided (Fidalgo, Alavi, & Amirian, 2014). In LR terms, concerns predominate with regard to "total score matching variable"(Li & Suen, 2013), the quality of "stepwise and systematic DIF analysis" (Hauger & Sireci, 2008), and potential misinterpretations leveled against the "magnitude of effect size" (Hidalgo & Lopez-Pina, 2004; Paek, 2012; Zumbo, 1999); therefore, a synopsis of these concerns together with how the present study is justified and dealt with within these concerns is in order.

According to Li and Suen (2013), generally within logistic regression method applied to DIF analysis, the total test score is considered as the matching variable. On the other hand, as Zhang (2006) warns us, using the total score as the matching variable may not work when the test is characterized by a multidimensional cognitive model. She proposed examinees' skill profile patterns as a criterion for matching. However, when many subskills are involved in a test, matching on profile patterns may not be practical (Li & Suen, 2013). In the context of PEET, six specialized subskills are involved in the test; therefore, PhD applicants could have as many as 64 (i.e. 2^6) specialized skill profiles. Given that the group sample size of the current DIF study was 602 for reference group and 397 for the focal group, matching PhD applicants on 64 (specialized) skill patterns was not practical. As such, the total scores from PEET were used as the matching variable in this study.

Another concern associated with DIF investigation is the step by step and systematic entering of variables into the LR equations. Researchers have argued that without ensuring this "stepwise procedure"(Hauger & Sireci, 2008), the probability of Type I and Type II errors would not be minimized (French & Maller, 2007; Navas-Ara & Gómez-Benito, 2002); and, accordingly, logical decisions would not be made with regard to fair testing and assessment (Hidalgo & Lopez-Pina, 2004).

In line with this concern, the present study applied a two-stage procedure for DIF analysis. Before embarking on the details of these stepwise procedures, some important points should be clarified. In LR approach, "the dichotomous logistic regression model is used to model the probability of correct response to the studied item as a function of observed test score (*X*), group membership (*G*), and the interaction of *X* and *G*" (Penfield & Camilli, 2007, p.139). For the present gender DIF study, the item response (1 for a correct response, and 0 for an incorrect response) was used as the dependent variable, with the independent variable being associated with grouping variable (1 = female /reference group; 0 = male/focal group), total scale score for each subject, and the interaction between total score and group membership. Moreover, in the present LR study, items flagged for DIF with negative directions were supposed to be in favor of reference group and those with positive values were claimed to be in favor of focal group.

As such, a two stage cycle of DIF analysis in LR was followed to compare models 1and 2 and, accordingly, to identify uniform and non-uniform DIF. In the first stage, the full form consisting of total score, gender and the interaction of the total score by gender were entered into the equation in model one. Then, in order to be certain that the interaction of total score by gender does not have any effect on the performance, this variable was removed from the equation in Model 2. As such, it was hypothesized If the −2 log-likelihood difference between Model 1 and Model 2 exceeds a $\chi 2$ value with 1 degree of freedom, potential (non-uniform) DIF would be possibly present. More information on details will be presented in the result section.

Full Model= Total score+ Gender +Interaction (T by G)

Reduced Model 1= Total score +Gender

In the second stage, as shown in the following equations, the total score and gender (Reduced Model 1) were entered as predictors. To assure that the differential performance on PEET subtests is not due to the effect of gender, gender was removed from the equation in Reduced Model 2. If the −2 log-likelihood difference between the two models is larger than a $\chi 2$ value with 1 degree of freedom, uniform DIF may be the result (for more information see the result section).

Reduced Model 1= Total score+ Gender

Reduced Model 2= Total score

All in all, the detailed procedure followed above was an attempt to minimize the effect of Type I and Type II errors. Nevertheless, simulation studies have reported that using a systematic LR approach without a measure of effect size could result in inflated Type I error (French & Maller, 2007; Hauger &Sireci, 2008; Jodoin & Gierl, 2001) and weaken the power of statistical tests (e.g. Cohen, 1988; Jodoin & Gierl, 2001; Hidalgo & Lopez-Pina (2004; Paek, 2012; Zumbo, 1999). Therefore, the present study followed the criteria established by Jodoin and Gierl (2001) to classify the items in terms of effect size. This criteria are presented in the following manner:

negligible or A-level DIF: R2 <0.035,

moderate or B-level DIF: Null hypothesis rejected AND 0.035 ≤ R2 <0.070,

large or C-level DIF: Null hypothesis rejected AND R2 ≥ 0.070.

However, in a recent study, Gómez-Benito, Hidalgo, and Zumbo (2013) recommended a different interpretation of effect size. They added that for the sake of accurate interpretation and appropriate decision making, a "blended decision rule"(Zumbo, 2008) including both the effect size and *p* value should be considered. Based on this recommendation, we applied both Nagelkerke R Square and Jodoin and Gierl's (2001) more conservative criteria to test the magnitude of gender DIF.

DIF analysis via IRT. The software used in the study was BILOG MG (Du Toit, 2003). BILOG MG has been introduced as the steadiest and most accurate software for the estimation of item parameters (Liu, Shu, & Jeng, 1998). This software has many applications for IRT including DIF analysis through 1-parameter IRT model. For the purpose of analyzing the data for DIF using IRT, at first the test was divided into different sections based on the content areas. As such, it was broken into seven sections. These sections were exactly the same sections separated by a title at the level of test design. Dividing a test into its subsections for DIF analysis is a procedure to bring about more accurate results and reduce the probability of type 1 error in which items are mistakenly marked for DIF (Clauser, & Mazor, 1998; Reeve, 2003).

Therefore, each subsection of the test was separately analyzed for DIF using 1-parameter IRT. To see whether each subsection suffered from DIF, a two-step analysis was used. In the first step, all the items of each subtest were analyzed in a single group as if they came from the same population, that is, male and female groups were considered to form one population. Then, the same data were analyzed in separate groups (for males and females) using the DIF model and under the null hypothesis of no DIF effects, the difference in the final log likelihood (labeled 2 LOG LIKLIHOOD in the output produced by Bilog MG) of the two stages was tested as x^2 with (n-1) (m-1) degree of freedom, where n is the number of items and m is the number of groups. When x^2 is significant, there is evidence that differential item functioning exists. The interpretation of this usually becomes clear when the item content is examined in relation to the direction of the estimated contrasts in the difficulty parameters. That is, when x^2 is significant, it indicates that DIF exists on that subtest. However, to see which items in that subtest suffer from DIF, analysis should be done at the level of items. Bilog MG provides the item difficulty and standard error for each item using 1-parameter IRT model. Items for which the threshold difference is roughly twice (1.96) or more the size of the standard error display DIF at the p = .05 level (Thissen, Steinberg, & Wainer, 1993). A lower threshold value (difficulty parameter) for a particular group means that the item is easier for them. That is, the negative or positive direction of the threshold differences indicates which particular subgroup is favored. In step one, if x^2 turns out to be insignificant for a particular subtest, it means there exists no DIF on that subtest. Therefore, there is no need to go through the second step and check the individual items for DIF.

Results

LR Results

As shown in Table 1, the −2 log-likelihood difference between the two models for each of the subtests was analyzed. The results showed that there is a −2 log-likelihood difference larger than the critical value of chi-square with 1 degree of freedom (i.e. $\chi 2$ (1, .05) = 3.84) for linguistics subtest, including two items (20 and 28), research methods consisting of one item(40), language assessment comprising one item(49), discourse including one item(95) and sociolinguistics comprising one item (99).

Table 2 presents the results for uniform DIF. In this case, six items exhibited the presence of uniform DIF with a difference larger than the critical value of chi-square with 1 degree of freedom (i.e. $\chi 2$ (1, .05) = 5.99). The subtests identified as showing DIF included linguistics, with two items (items 23& 26), research with one item (item36), testing with one item (item 46), SLA with one item (item 78), and sociolinguistics with one item (item 97). As such, the information reported in Tables 1 and 2 indicate that 12% (12 items) of the whole test were identified as showing DIF with equal numbers of items showing uniform (six items) and non-uniform DIF (six items).

Table 1
Summary of −2 Log-likelihood Differences of Stage 1 Analysis

Subtests	Items	−2 log-likelihood of Full Model	−2 log-likelihood of Reduced Model 1	−2log-likelihood difference between Full Model & Reduced Model 1
Linguistics	20	1069.532	1074.020	4.488*
	23	997.418	997.574	0.156
	26	665.077	666.660	1.583
	28	1060.263	1064.453	4.19*
Research Methods	36	1163.156	1163.168	0.012
	40	907.302	911.465	4.163*
Language	46	1024.212	1024.258	.046
Assessment	49	240.757	246.543	5.286*
SLA	78	684.944	685.209	0.265
Discourse	95	830.816	839.895	9.079*
	97	691.706	692.249	0.543
Sociolinguistics	99	582.733	587.164	4.431*

Note: * Larger than the critical value of $\chi 2$ (1, .05) = 3.84.

Table 2
Summary of −2 Log-likelihood Differences of Stage 2 Analysis

Subtests	Items	−2 log-likelihood of Reduced Model 1	−2 log-likelihood of Reduced Model 2	−2log-likelihood difference between Reduced Models 1 & 2
Linguistics	20	1074.020	1074.125	0.105
	23	997.574	1005.371	7.797*
	26	666.660	675.308	8.648*
	28	1064.453	1064.487	0.034
Research Methods	36	1163.168	1170.468	7.3*
	40	911.465	911.511	0.046
Language	46	1024.258	1030.522	6.264*
Assessment	49	246.043	248.292	2.249
SLA	78	685.209	694.650	9.441*
Discourse	95	839.895	844.243	4.348
Sociolinguistics	97	692.249	703.131	10.882*
	99	587.164	588.319	1.155

Note: * Larger than the critical value of $\chi 2$ (2, .05) =5.99

The overall results of LR DIF are summarized in Table 3. Worthy of note is that only items flagged for significant DIF values at 0.05 level of significance are included in the table. As Table 3 displays, the obtained R2 values report that gender DIF is distributed equally between uniform and non-uniform DIF on different subtests of PEET. Of the 12 items identified as showing DIF, six items

have a significance value larger than the critical value of chi-square with 1 degree of freedom (i.e. $\chi 2$ (1, .05) = 5.99), that is they are uniform. Equally, six items have significance values larger than the critical value of chi-square with 1 degree of freedom (i.e. $\chi 2$ (1, .05) = 3.84); that is they are non-uniform. Out of 12 items, 4 items were detected in the linguistics section, two in the research subtest, two in Testing, one in SLA, one in Discourse and finally two items in Sociolinguistics. With regard to DIF effect size, the present study followed a "blended decision rule"(Zumbo, 2008) including both the effect size and *p* value. Likewise, it was observed that all obtained R2 values manifested a negligible DIF magnitude (category A); that is, they were smaller than .035 and .05.

Table 3

Uniform, Non-uniform, Total R2 Effect Sizes, and the Chi-squared Test Results

			R2 **effect size**				Category
Item	subtest	Favored	UDIF	NUDIF	DIF	$\varkappa 2$	
20	L	M	.....	.005	.005	4.488	A
23	L	F	.008	.....	.008	7.953	A
26	L	M	.012	.....	.012	10.231	A
28	L	F	.....	.005	.005	4.19	A
36	R	F	.008	.....	.008	7.312	A
40	R	M	.....	.005	.005	4.163	A
46	T	M	.007	.....	.007	6.31	A
49	T	F	.....	0.019	.019	12.821	A
78	SL	F	.016	.....	.016	9.706	A
95	D	F	.....	.000	.000	22.506	A
97	S	M	.015	.....	.015	11.425	A
99	S	M	.....	.006	.006	4.431	A

Notes. **p* <.05; L= Linguistics; R= Research; T= Testing; SL = SLA; D= Discourse; S= Sociolinguistics; M= Male; F= Female; A = Negligible DIF

IRT Results

The results of DIF analysis based on the IRT model indicated that overall three subtests (Skills, Discourse & Socio, and Research methods) did not suffer from DIF, while four subtests, namely, Teaching Methods, Linguistics, Language Testing and SLA were flagged for DIF. Overall seven items were flagged with DIF, 5 items favoring females and 2 items favoring males, the details of which will be explained below.

In each section, the difficulty differences between the contrasting groups, called group threshold differences, and the standard error of measurement are provided for items flagged with DIF. As displayed in Table 4, the threshold difference for these items are roughly twice (1.96) or more the size of the standard error at the *p* = .05 level (Thissen, et al, 1993). The only parameter to be attended to in this program was the difficulty value (b) and therefore the lower threshold value for a particular group means that the item was easier for them. That is, the negative or positive direction of the threshold differences indicates which particular subgroup was favored.

Table 4
Group Threshold Differences for Items Indicating DIF

Subtest	ITEM	GROUP 2 - 1
Teaching Methods	2	-0.539
		0.251*
Linguistics	17	-0.854
		0.258*
	22	-0.880
		0.252*
	30	-0.712
		0.347*
Research methods		NO DIF
Language testing & Assessment	32	-0.598
		0.228*
	33	0.479
		0.193*
Skills		NO DIF
SLA	82	0.828
		0.337*
Discourse and Socio	NO DIF	

*Standard Error

As demonstrated in Table 4, only one item displayed DIF in the teaching method subtest (item 2) with a threshold difference value of -0.539 being roughly twice as much as the standard error (0.251). The negative threshold difference reported for this item indicates that this item is easier for (more in favor of) females. The second subtest displaying DIF in the IRT One-parameter analysis was Linguistics. As shown in Table 4, three items (items 17, 22, 30) were flagged with DIF, all being easier for females. Item 17 showed a threshold difference of -0.854 being twice as much as the standard error (0.258). For item 22, the threshold difference was reported to be -0.880, again being twice as much as the standard error (0.252). Like those of items 17 and 22, the threshold value reported for item 30 (-0.712) was twice as much as the reported standard error (0.347).The third subtest showing DIF in the IRT analysis was language testing and assessment. As displayed in table 15, two items were flagged with DIF in this subtest (items 32 & 33). With a negative threshold value of -0.598, and a standard error of 0.228, item 32 was in favor of females. However, the positive threshold value of 0.479 reported for item 33 shows that this item is easier for males. The last subtest displaying DIF in the IRT analysis was SLA. Only one item (item82) was flagged with DIF. As demonstrated in Table 4, the threshold value reported for this item is 0.828, being in favor of or easier for males. All in all, it can be said that, seven items were flagged with DIF; five items favoring females and 2 items favoring males.

The Comparison between LR and IRT Results

Table 5 summarizes the overall results of gender DIF for both LR and 1-p IRT. It is reported that12 items were flagged with DIF in LR method, while IRT-one parameter method showed 7 items indicating DIF. This finding is in line with the dominant view in DIF literature that LR detects more DIF items in comparison to other techniques due to its power of detecting both UDIF and NUDIF (Hidalgo, & López-Pina, 2004; Rogers & Swaminathan, 1993). Speaking metaphorically, scholars proclaim that LR feels free to accuse an item of displaying DIF (Jodoin & Gierl, 2001, Rogers & Swaminathan 1993).

It was also shown that there was inconsistency in the results of the two methods in terms of detecting gender DIF in the type of items across the subtests of PEET. As far as different subtests of this test are concerned, LR detected 4 DIF items in Linguistics, two items in Research, two items in language testing, one in SLA and three items in Discourse and Sociolinguistics with no items detected for DIF in Teaching Methods and Skills sections, whereas IRT flagged 2 DIF items in Teaching methods, three items in linguistics, two items in language testing and one in SLA with no items identified as showing DIF in Research Methods, Skills, Discourse and Sociolinguistics. Linguistics, Language Testing, and SLA were among the sections the items of which were identified as showing DIF by both methods. Both methods detected no gender DIF items for the skills section and no individual items were jointly detected by either methods.

As far as the magnitude of DIF is concerned, however, it was found that all DIF items detected by LR method displayed a negligible or type-A effect size while IRT flagged 7 items for DIF.

Table 5
Items Flagged for DIF in both LR and IRT Models

Subtests	LR	IRT	Both LR and IRT
Teaching Methods	.....	2	.....
Linguistics	20, 23, 26, 28	17, 22, 30	.....
Research	36, 40	.....	.....
Language testing	46, 49	32, 33	.....
Skills	.....	.....	.....
SLA	78	82	.....
Discourse & Socio	95, 97, 99	.....	.....

Discussion, Conclusions and Implications

This study examined gender DIF on the PhD Entrance Exam of TEFL (PEET) in Iran, using both LR and 1-p IRT models. The results indicated that group membership can significantly affect the performance on the PEET as illuminated through DIF analysis, though, the DIF results differed a great deal depending on the analytic method used; the LR procedure identified a larger number of DIF items than did the 1-p IRT procedure. However, in LR the DIF cancellation occurred (there was equal number of DIF items for males and females), while most of the DIF items identified in 1-p IRT were in favor of females.

In LR terms, these findings highlight the existence of gender DIF. Although the PEET showed as many as 12 DIF items, of particular interest is the finding that the number of DIF items for males and females was equal, indicating that DIF items might balance out each other in the test level analysis (Drasgow, 1987; Takala & Kaftandjieva, 2000), what Sireci and Rios (2013) refer to as, DIF "cancellation". In this regard, the findings of the present study are partially in keeping with those of Park's (2008) study, though in a different testing application context. Since no specific study has been carried out across subject matter test of TEFL, replication studies are needed to investigate gender DIF in this context.

With regard to the magnitude of gender DIF, it was found that all the items flagged for DIF were classified as negligible (category A), that is, they were smaller than .035 and .05. Since effect size is somehow influenced by sample group size, the interpretation on the magnitude of DIF should be treated with caution. For the present study, test score data of a total of 999 applicants consisting of 397 males and 602 females were subjected to gender DIF investigation. Some scholars justify a sample size of 200 per group as sufficient to add power to LR results and consequently avoid inflated Type I error (Güler& Penfield, 2009; Mazor, Clauser, & Hambleton, 1992; Paek & Wilson, 2011; Zumbo, 1999). Accordingly, in the present LR study the magnitude of DIF interpretation may be less in doubt in this regard; nevertheless, Jodoin and Gierl (2001) found that large differences in sample size across groups may impede DIF detection. Moreover, Sireci and Rios (2013) state that:

> In some cases, the sample sizes for the reference group are much larger than for the focal group. When this occurs, multiple random samples can be taken from the reference group, and the analyses can be replicated using the same group of focal group examinees" (p. 183).

It seems this may hold true for the present study, since the size of the female group (602) was considerably larger than the size of the male group (397). This inequality may cast some doubts on the interpretation of the magnitude of DIF in the present study. Surprisingly, had we equalized the sample sizes across groups, we may not have added to the certainty of the DIF interpretation, this being in line with Gomez's (2008) finding in which case the reference group size did not affect the Type I error of the LR procedure and the highest rates of false positives for the LR procedure were observed with equal groups of 1500 examinees. Likewise, with these perplexing findings in the literature which well indicate the complexity of DIF conceptualization and the significance of appropriate DIF method selection, more studies with different statistical DIF detection procedures are warranted to check the influence of equal or unequal comparison group sample sizes on the magnitude of DIF interpretation.

Another important finding emerging from the present study is that the number of nonuniform gender DIF items was equal to uniform DIF (see Table 3). Non-uniform group effect exists when the direction of the relative advantage of one group over another is changed at some point on the ability scale being not systematic across the entire ability continuum, that is, there is an interaction between group membership and ability level (Mellenbergh, 1982). As far as the present study is concerned, the interaction of total score by gender was a good predictor of differential performance. This finding is partially in contrast with the literature in which the predominance of uniform DIF is the most typical situation (e.g., Ahmadi & Jalli, 2014; Amirian, Alavi & Fidalgo, 2014; Breland& Lee, 2007; Rezaee & Shabani, 2010). Many a language teacher and SLA researcher have questioned and challenged the parameter of gender as a fixed, binary variable that is often embraced in gender research in language learning (e.g. Ehrlich, 1997; Sunderland, 2000). They claim that rather than being a fixed, biological variable, gender is predominantly a socially constructed variable within specific cultural and situational contexts (Davis & Skilton-Sylvester, 2004). The sample groups participating in the present study were all PhD applicants with different socioeconomic backgrounds. Likewise, there might be different effects of gender and its interaction with the ability level for applicants with different socioeconomic statuses. Therefore, in this context

looking at the notion of gender as a "fixed biological variable" (Davis & Skilton-Sylvester, 2004) and making any forceful interpretation on its possible effect would be in doubt. Put another way, had we taken these variables into account, we might have come to different results. As such, a comprehensive study is needed to classify the examinees into several major cultural, national, and educational subgroups and conduct a separate gender DIF study within each of these subgroups (Breland & Lee, 2007). This can be done within a context similar to the one in the present study or other similar research investigating gender DIF. Moreover, this study applied statistical analyses to identify DIF. Further research can explore whether bias reviewers can identify test items flagged for DIF without statistical data (Engelhard, 1990).

As mentioned before, the second DIF detection method used in this study was 1-p IRT model. It was shown that as many as 7 items were flagged for DIF using this model. The findings predominantly highlight female over-performance on different subtests of PEET. Of the total of 7 items IRT model detected as showing DIF, five were in favor of females. Put another way, about 71% of the items flagged for DIF were differentially easier for females. One source of explanation may rely on what some scholars agree upon and accept as a controversial idea of heated debate in the field of second language acquisition (SLA); the widespread belief of female superiority in language learning (Breland & Lee, 2007; Davis & Skilton-Sylvester, 2004). Some theorists and researchers in the field of second language acquisition (SLA) have claimed that females are better language learners in general and, accordingly, have superior linguistic ability overall (Ehrman & Oxford, 1988), although previous research has produced mixed results (Breland & Lee, 2007). Unfortunately, no specific study has been conducted to investigate this issue in subject matter tests of TEFL; therefore, strongly supporting the general conclusion of female superiority in PEET test performance is controversial or at least difficult. As such, a replication study using another sample (with another administration of PEET) of the same population of PhD applicants is warranted to determine whether this finding is simply a function of group differential performance or test impact.

Still another important finding emerging from the comparison of the results of LR and IRT models is that little consistency was observed between LR and IRT findings in DIF detection, though both models identified gender DIF in three sections in common. This finding is partially in keeping with Ahmadi and Jallili's (2014) study, reporting a low level of overlap between the results of the two methods. One source of explanation may be the model of IRT used to detect gender DIF across PEET. Likewise, this study merits further investigations, using two or three parameter IRT in DIF detection.

As regards the direction of the gender DIF, of a total of 12 items, LR detected equal number of items favoring males and females (6 items favored males and six items favored females); therefore, DIF cancellation might have occurred at least at the level of number of DIF items favoring each group. However, DIF cancellation is a tricky issue that depends on the number of items indicating DIF as well as the magnitude of DIF (see Pae, & Park 2006; Zumbo, 2003, for a relevant discussion). On the other hand, IRT detected 7 items of which five items favored females and two favored males. One possible explanation for this inconsistency is that both LR and IRT are sensitive to the nature of the data and the group sample size and both methods treat these factors differently; nonetheless, we see that in the present study, the same data and group sample size have been used with both models. As Zumbo (1999) puts it, IRT methods are more at ease with large sample sizes and LR models may not show reliable results if Bonferroni correction test could not be applied (Alavi & Karami, 2010; Runnels, 2013; Thompson, 2006), and multiple random samples could not be taken from the reference group (Sireci & Rios, 2013), especially when we have unequal group sample sizes. In this way, lower number of items can be identified as showing DIF (Ahmadi & Jallili's, 2014). As such, a more thorough study is urgent to attend to the unequal sample sizes across groups when using LR and IRT models.

The findings of the present study may have several important implications for test developers, test takers, researchers and TEFL teachers at post graduate levels. As a high-stakes test, the PhD Entrance Exam may have a great impact on the instructional practice of university teachers at MA and PhD levels. They may direct their teaching toward the successful performance of their students on this test. They may also decide to include the topics covered in this test in their course syllabi. For instance, differential item analysis of this test showed that a number of items in different areas of TEFL were likely in favor of females. Based on this finding, instructors may provide additional sustained help on these areas and encourage PhD candidates to pay due attention to these specific areas.

Based on the findings of the present study, it was shown that some subtests such as Linguistics were predominantly differentially in favor of females. Considering the personal and social ramifications of PEET and other similar gate-keeping instruments, test developers and policymakers need to evaluate the present and other similar gender DIF findings when making decisions about PhD applicants. For example, after reviewing the subtests the items of which were identified as DIF by both LR and IRT, test developers can decide to change the content of those subtests. Selection committee and bias reviewers appointed by NOET or MSRT can also make a thoughtful evaluation on the content of DIF items and remove those which are reported to be biased. In the context of higher education in Iran, this can bring the task of decision making to a sort of even-handedness and promote the validity of interpretations from educational and psychological assessments. Of course, decisions about items indicating DIF (omitting the items, modifying the items, or simply ignoring the items, e.g., because the items in favor of different sub-groups may cancel each other out) are best served when the causes of DIF are known. The present study seems to be the first one of its type conducted on the subject-matter tests within the field of TEFL, so not much could be recommended in this regard before further studies focus on illuminating the causes of DIF on such tests. Until then, the most logical implication would be to omit such items from the tests, especially when it comes to high-stakes tests, to avoid unintended detrimental consequences of using the test results. This, of course, further highlights the care that test developers should take at the time of test design in order to avoid DIF as far as possible.

Given that the results of the present study revealed some gender DIF items, top tier decision makers (like NOET & MSRT) can take these findings into serious consideration and exercise care in fair test practice by dedicating effort to more unbiased test development and decision making. Before introducing their tests, test developers can make sure the tests they have developed enjoy quality control and quality assurance by subjecting them to external review or by receiving adequate training in psychometrics (Zandi, Kaivanpanah, & Alavi, 2014). Accordingly, test takers (e.g., PhD applicants) may be more appropriately evaluated based on their true knowledge or language ability. Economically at least, this would be of great benefit to PhD applicants (true positives), since otherwise (introducing false negatives) they would have to apply for and register in non-financial universities and pay huge sums of money to universities as their tuition or they would be urged to stay at home and risk one more year-long preparation. This may also create some psychological and social problems for them.

Further, in line with the previous studies (e.g., Ahmadi, & Jallili, 2014), though on a different testing application context, the present study revealed little consistency between LR and IRT findings in DIF detection. Testing researchers can benefit from this finding and can make a replication study to see whether the same findings are repeated.

Moreover, in the present study, the results from LR model revealed that gender DIF items may cancel each other out at the item level. Worthy of note is that the studies of DIF to date have not shown whether in the test level analysis the accumulation of DIF items cancel each other out (Park, 2008). Using the information about the DIF cancellation provided by the present gender DIF

study, and considering what Park (2008) introduces as a gap, researchers can extend this study and investigate DIF cancellation at the total test score level.

The current study may also contribute to the DIF literature by providing information about DIF across a subject matter test (TEFL) with an Iranian sample. To the best of our knowledge, no specific gender DIF study on TEFL has been carried out in national and international testing applications and it is not clear whether DIF findings on such tests may be common across nationalities; thus, the findings of this study may provide valuable information by helping to bridge this research gap in DIF studies. Furthermore, considering the paucity of DIF research carried out across non-English major tests in the context of higher education, and relying on what Douglas (2014) emphasizes as "basing language training and assessment on a language for specific purposes foundation"(Douglas, 2014, p.2), researchers can benefit from the present findings and can make replication studies on gender DIF across other subject matter tests beyond TEFL.

There is a limitation of this study (i.e., the unequal group sample size) that should be addressed in future research. Given that in this study the size of the reference group was almost twice as much as the size for the focal group, this might have polluted the validity of DIF interpretation. Likewise, potential studies that apply Bonferroni correction test (Alavi & Karami, 2010; Runnels, 2013; Thompson, 2006) and use multiple random samples (Sireci & Rios,2013) from reference group with focal group as being fixed are welcomed in later studies.

References

Ahmadi, A., & Jalili, T. (2014). A confirmatory study of Differential Item Functioning on EFL reading comprehension. *Applied Research on English Language*, *3*(6), 55-68.

Alavi, S. M., & Karami, H. (2010). Differential item functioning and ad hoc interpretations. *TELL, 4*(1), 1-18.

Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, *4*(2), 187-203.

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

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

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

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two meta-analyses. *Journal of Research in Science Teaching, 26,* 141–169.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27,* 165–174.

Breland, H., & Lee, Y-W. (2007). Investigating uniform and non-uniform gender DIF incomputer-based ESL writing assessment. *Applied Measurement in Education*, *20*, 377–403.

Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal, 34*, 297–331.

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

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–47.

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

Cole, N. S. (1997). *The ETS gender study: How males and females perform in educational settings.* Princeton, NJ: Educational Testing Service.

Davis, K. A., & Skilton-Sylvester, E. (2004). Looking back, taking stock, moving forward: Investigating gender in TESOL. *TESOL Quarterly, 38*(3), 381–404.

Douglas, D. (2014). Nobody seems to speak English here today: Enhancing assessment and training in aviation English. *Iranian Journal of Language Teaching Research 2*(2), 1-12

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *The Journal of Applied Psychology, 72,* 19–29.

Ehrlich, S. (1997). Gender as social practice. Implications for second language acquisition. *Studies in Second Language Acquisition, 19*, 421–446.

Ehrman, M., & Oxford, R. (1988). Effects of sex differences, career choice, and psychological type on adult language learning strategies. *Modern Language Journal*, *72*, 253–265.

Engelhard, G. (1990). Gender differences in performance on mathematics items: Evidence from the United States and Thailand. *Contemporary Educational Psychology, 15*, 13–26.

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

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with Logistic Regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*, 373–393.

Geranpayeh, A., & Kunnan, A. J. (2007) Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), 190-222.

Gierl, M, Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality based DIF analysis. *Journal of Educational Measurement, 40*(4), 281–306.

Gómez-Benito, J., Hidalgo, M. D., & Zumbo, B. D. (2013). Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous items. *Educational and Psychological Measurement*, *73*(5), 875-897.

Güler, N., & Penfield, R. D. (2009). A comparison of logistic regression and contingency table methods for simultaneous detection of uniform and non-uniform DIF. *Journal of Educational Measurement*, *46*, 314–329.

Haeok, K. (1990). A longitudinal study of sex-related bias in mathematics subtests of the California Achievement Test. *Applied Measurement in Education*, *3*, 275–284.

Halpern, D. (1992). *Sex differences in cognitive abilities*. Hillside, NJ: Lawrence Erlbaum.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*, 211–235.

Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29 (2),* 163-180.

Harris, A., & Carlton, S. (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. *Applied Measurement in Education, 6,* 137–151.

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

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel–Haenszel procedures. *Educational and Psychological Measurement*, *64*, 903–915.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hyde, J., & Linn, M., (1988).Gender differences in verbal activity: A meta-analysis. *Psychological Bulletin, 104*, 53–69.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating power and type I error rates using an effect size with the logistic regression procedure for DIF. *Applied Measurement in Education, 14,* 329–349.

Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society, 26,* 352–366.

Kiany, R., Shayestefar, P., Ghafar Samar, R., & Akbari, R. (2013). High-rank stakeholders' perspectives on high-stakes University entrance examinations reform: priorities and problems. *Higher Education, 65*, 325–340.

Kim, Y. H., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*(4), 825-865.

Li, H., & Suen, H. K (2013). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing*, *30*, 273–298.

Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology*, *35*, 299–314.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443–451.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105–107.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289–304.

Näsström, G. (2004). *Differential item functioning for items in the Swedish national test in mathematics course*. Retrieved from http://www. Vxu.se/msi/picme1o/L2ng.PDF.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, *18*, 9–15.

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

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

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

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

Paek, I. (2012). A note on three statistical tests in the logistic regression DIF procedure. *Journal of Educational Measurement*, *49*, 121–126.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the Marginal Maximum Likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*, 1023–1046.

Park, G.-P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly, 42*(1), 115–122.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay, & C.R. Rao (Eds.), *Handbook of statistics, Volume 26 Psychometrics* (pp.125-167). New York: Elsevier.

Plake, B. S. (1980a). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40*, 397–404.

Plake, B. S. (1980b). An investigation of the Iowa Tests of Basic Skills for sex bias: A developmental look. *Psychology in the Schools*, *17*, 47–52.

Reeve, B. B. (2003). An introduction to modern measurement theory. Retrieved from http://appliedresearch.cancer.gov/areas/cognitive /immt.pdf.

Rezaee, A., & Shabani, E. (2010). Gender differential item functioning analysis of the University of Tehran English Proficiency Test. *Pazhuhesh-e Zabanha-ye Khareji*, *56*, 89–108.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel–Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105–116.

Runnels, J. (2013). Measuring differential item and test functioning across academic disciplines. *Language Testing in Asia*, *3*(9), doi:10.1186/2229-0443-3-9.

Ryan, K., & Bachman, L.F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12–29.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*, 170–187.

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

Stoneberg, B. D. (2004). *A study of gender-based and ethnic-based differential item functioning (DIF) in the spring 2003 Idaho Standards Achievement Tests. Applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel Chi Square Test*. Paper for EDMS 889 Measurement-Statistics Practicum, University of Maryland, College Park. Retrieved from http://files.eric.ed.gov/fulltext/ED483777.pdf

Sunderland, J. (2000). Issues of language and gender in second and foreign language education. *Language Teaching, 33*, 203–223.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*, 323–40.

Thissen, D., Steinberg, L., & Wainer, H., (1993). Detection of Differential Item Functioning using the parameters of item response models. In Holland, P.W., & Wainer, H. (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associate, Hillsdale, NJ.

Thompson, B (2006). *Foundations of behavioral statistics: An insight-based approach*. London: The Guilford Press.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing, 22*, 211–234.

Young, D. J., & Fraser, B. J. (1994). Gender differences in science achievement: Do school effects make a difference? *Journal of Research in Science Teaching, 31,* 857–871.

Zandi, H., Kaivanpanah, SH., & Alavi, S.M. (2014). The effect of test specifications review on improving the quality of a test. *Iranian Journal of Language Teaching Research 2(1), 1-14.*

Zenisky, A., Hambleton, R., & Robin, F. (2003). Detection of differential item functioning in large scale state tests: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*, 51–64.

Zhang, W. (2006). Detecting differential item functioning using the DINA model (Unpublished doctoral dissertation). University of North Carolina at Greensboro, Greensboro, NC.

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.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*(2), 136-147.

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B. D. (2008, July). Statistical methods for investigating item bias in self-report measures. *Florence Lectures on DIF and Item Bias*. Lectures Conducted from Universita degli Studi di Firenze, Florence, Italy.

Alireza Ahmadi is an Associate Professor in Teaching English as a Foreign Language (TEFL) at Shiraz University, Shiraz, Iran. He received his PhD from the University of Isfahan, Iran in 2008. He has published about 20 articles in scholarly journals. His research centers on Second Language Assessment and Language Teaching.

Ali Darabi Bazvand is a PhD candidate in the Department of Foreign Languages and Linguistics at Shiraz University. He has published several papers in scholarly journals. His areas of interest include Second Language Assessment and Teacher Education.