Chabahar Maritime University
Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, 2020, 9(3) (Previously Published under the Title: Maritime English Journal)

Does Field of Study Matter in Academic Performance: Differential Item Functioning Analysis of a High-Stakes Test Using One-Parameter and Two-Parameter Item Response Theory Models

[1]Masoomeh Estaji*
[2]Kiyana Zhaleh

IJEAP- 2005-1551
Received: 2020-05-29 Accepted: 2020-07-11 Published: 2020-07-13

Abstract

To ensure test fairness and validity in high-stakes tests, in the absence of sufficient evidence for identifying various sources of Differential Item Functioning (DIF), the present study examined the effect of field of study on the reading section of the English subtest of Iranian University Entrance Examination (IUEE)for MA in English majors. 1-parameter and 2-parameter logistic Item Response Theory (IRT) models were employed to investigate DIF for a sample of 3588 applicants sitting for the test in 2017. For data analysis, the difR package developed by Magis, Beland, Tuerlinckx, and De Boeck (2010) was utilized. The 1-parameter DIF analysis results indicated that out of the 20 items of the reading section, only three items showed DIF toward the examinees based on their field of study. However, 2-parameter DIF analysis results demonstrated that all items of the reading section presented DIF toward the examinees. Furthermore, the least discrimination toward the non-English group was observed in the findings of the 2-p IRT model. Hence, the 2-p IRT model was found more accurate than the 1-p IRT model as it could identify more DIF items, and the reading comprehension subtest of IUEE was biased toward examinees from different fields of study. Based on the results, to identify and remove various sources of potential DIF existing in the tests and produce test items which are void of any bias in terms of academic background, the use of IRT models is required; although their level of precision varies.

Keywords: Differential Item Functioning (DIF), difR Package, Field of Study, Iranian University Entrance Examination (IUEE), Item Response Theory (IRT), Reading Comprehension

1. Introduction

Test fairness, as an issue of great significance in the area of educational measurement and particularly, language testing, has to do with the consideration of test design quality, test administration and scoring, test content coverage and relevance, and test construct validity (Alavi & Bordbar, 2017). Furthermore, it aims at providing equal opportunity of learning and access to testing to all examinees as well as identifying test items that assess the ability under measurement, not the factors introducing construct-irrelevant variance in test scores (Shohamy, 2000). Consequently, when the test items measure test-takers' characteristics other than the ability under investigation, they (dis)advantage specific examinees groups and introduce bias in testing (McNamara & Roever, 2006). To be more specific, a test or an item is found to be biased if the test takers from dissimilar groups with equal ability levels have a differential probability of providing correct responses to the test items (Camilli & Shepard, 1994). Such test-takers' characteristics, which might bring about systematic construct-irrelevant variance in test scores and, therefore, cause test bias are, to name but a few, ethnicity, gender, race, first language, and socioeconomic and academic backgrounds (McNamara & Roever, 2006).

While test fairness is recognized as a key component of constructing valid tests, test bias is regarded as a construct-irrelevant variance threatening test validity (Hope, Adamson, McManus, Chis, & Elder, 2018). Test bias pertains to the overall description of test situations where construct-

[1] Associate Professor of Applied Linguistics (Corresponding Author), mestaji74@gmail.com, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran.

[2] PhD Student in TEFL, k_zhaleh97@atu.ac.ir, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran.

irrelevant group characteristics impact test scores (Kane & Bridgeman, 2017). In order to identify test bias, a technical term named Differential Item Functioning (DIF) was introduced, which relates to the investigation of test items functioning differentially within various groups of examinees (Martinková et al., 2017). Through DIF analysis, one can provide construct validity evidence through the internal structure of the test. Furthermore, when items potentially biased against particular subgroups of examinees have been identified, they can be screened out by item writers to increase test validity (Gómez-Benito, Sireci, Padilla, Hidalgo, & Benítez, 2018).

Most of the DIF studies to date have been conducted on large-scale assessment situations, and consequently, their findings are generalized to large groups, not particular examinees (McNamara & Roever, 2006). The justification for prevalent investigation of DIF on high-stakes tests is that promoting fairness in high stakes tests is an important priority as decisions based on the results of such tests impact people's lives (De Beer, 2004). Moreover, the consequences of poor performance on these tests are grave as applicants should spend more time, money, and effort to prepare for the tests in the following years. Hence, it is crucial to evaluate these tests for their psychometric properties of reliability and validity (Pae, 2011).

As a high stakes test, Iranian University Entrance Examination (IUEE) is administered each year among a large number of applicants competing to enter universities. Although a myriad of empirical studies has examined various high-stakes tests for identifying items showing DIF (e.g., Ahmadi & Jalili, 2014; Chen & Henning, 1985; Fidalgo, Alavi, & Amirian, 2014; Geranpayeh & Kunnan, 2007; Hope at al., 2018; Li, Cohen, & Ibarra, 2004; Oliveri, Lawless, Robin, & Bridgeman, 2017; Pae, 2004), few studies have analyzed IUEE items for DIF identification, most of which have focused on investigating the likelihood of gender-DIF among IUEE items, and scant attention has been paid to the field of study as a source of DIF in this test (e.g., Ahmadi & Darabi Bazvand, 2016; Alavi & Bordbar, 2017; Barati & Ahmadi, 2010; Ravand, Firoozi, & Rohani, 2019; Salehi & Tayebi, 2011). In the absence of sufficient evidence for identifying various sources of DIF, the present study endeavored to investigate the role of test-takers' field of study as a source of test bias in IUEE through examining 1-p and 2-p logistic Item Response Theory (IRT) models, using the difR package developed by Magis et al. (2010).

2. Literature Review

2.1. Theoretical Framework

Differential Item Functioning is of two types of uniform DIF and non-uniform DIF. More particularly, Uniform DIF happens "when a group performs better than another group on all ability levels" (Karami, 2012, p. 60), group membership does not interact with level of ability. However, non-uniform DIF happens in situations that "members of one group are favored up to a level on the ability scale and from that point on the relationship is reversed", and an interaction exists between group membership and level of ability (Karami, 2012, p. 60). There are at least two groups when running DIF analyses, classified as either focal or reference groups. The focal group relates to the minority group while the reference group pertains to the majority group (Cuevas & Cervantes, 2012). Regardless of the method used to detect DIF, the focal group's item responses are compared to those of the reference group in order to identify items bringing about different performance of the two groups.

DIF detection can be done through various statistical methods (Millsap & Everson, 1993) such as logistic regression (Swaminathan & Rogers, 1990), standardization approaches (Dorans & Kulick, 1986), simultaneous item bias test (Shealy & Stout, 1993), Mantel–Haenszel statistics (Holland & Thayer, 1988), and IRT (Hambleton, Swaminathan, & Rogers, 1991). Regardless of what method is employed, each item is usually analyzed separately for DIF identification and subsequently classified according to the DIF present (Wainer, 1995). In IRT-based methods, the examinees are matched according to their level of ability, i.e., the latent trait (θ). One crucial element of IRT models is the Item Characteristic Curve (ICC) which is an S-shaped curve indicating the association of examinee's ability level, as indicated on the horizontal axis, as well as his/her likelihood of answering correctly to a test item, as demonstrated on the vertical axis

(DeMars, 2010). Typically, the ability is computed in logit units ranging from –4 to +4, with 0 value showing an ability level by which the examinee has an equal probability of (in)correctly answering an item with an average difficulty level (McNamara & Roever, 2006).

In order to describe the ICC shape, IRT utilizes three parameters, namely item difficulty, item discrimination, and guessing factor. IRT has three models, including one, two, and three logistic IRT models, depending on how many of these parameters are considered in exploring the association of the ability level with item response patterns (McNamara & Roever, 2006). In the 1-p IRT model or/and the Rasch model, the level of ability is related to item difficulty to model the likelihood of providing correct naswer to the test item (Ockey, 2007). As to the items differing on the difficulty parameter, the probability of providing a correct response is higher for a low-difficulty item compared to a high-difficulty item (McNamara & Roever, 2006). In the 2-p IRT model, besides item difficulty, item discrimination, displayed in the slope of the ICC, is also added to the model (Ockey, 2007). Finally, the 3-p IRT model inserts a guessing factor besides the parameters of item discrimination and item difficulty (Ockey, 2007).

There are various assumptions underlying all the three models of IRT, namely uni-dimensionality, local independence, model-to-data fit, large sample size, and the certainty that test-takers' responses are a true indicator of their ability level. IRT models are advantageous since they provide sample-independent item-parameter and test-independent test takers' ability level estimation. In IRT models, we also have item economy, item banking, multiple measures of reliability, and the reconciliation of criterion- and norm-referenced testing.

2.2. Empirical Studies

Many empirical research investigations have run DIF analysis with the aim of detecting test bias in various high-stakes tests. To name a few, Chen and Henning (1985) examined the presence of cultural and linguistic bias in the English Second Language Placement Examination held at California University. Their findings indicated that from the total of 150 items in the test, only four vocabulary items were identified as biased based on the criteria of the study. Linn, De Benedicts, Delucchi, Harris, and Stage (1987), looking for the potentiality of gender-DIF among National Assessment of Educational Progress Science items, reported that items related to masculinity or physical science content were biased toward female examinees. Hale (1988) analyzed examinee's responses to the reading section of TOEFL to examine if test takers' academic discipline led them to function differently on this test. The outcomes of this study showed that examinees from physical sciences and social sciences performed significantly better on passages pertaining to their own field of study than passages related to other majors. Takala and Kaftandjieva (2000), who investigated the likelihood of DIF in the vocabulary subtest of the Finnish Foreign Language Certificate Examination, found that on the whole test was not biased toward neither males nor females.

Geranpayeh and Kunnan (2007) ran DIF analysis on the listening section of the Certificate in Advanced English examination to probe whether the test showed bias toward examinees from diverse groups of age. Findings obtained from content and statistical analyses revealed DIF in some of the test times. In 2015, Song, Cheng, and Klinger examined DIF in Graduate School Entrance English Examination in China with respect to test taker's academic background and gender through using SIBTEST and content analysis. Their results revealed poor reliability for many of the test items, leading to their inability to distinguish between high- and low-test performance. Moreover, Oliveri et al. (2017) detected possible sources of DIF in the quantitative reasoning and verbal reasoning sections of all eight forms of a GRE exam. The outcomes suggested that some of the test items were biased toward examinees based on their citizenship group. In a recent study, Hope et al. (2018) investigated DIF likelihood in a high-stakes postgraduate knowledge-based assessment. The findings demonstrated that from 2773 items in the test, only eight items showed significant DIF.

In Iran, Barati and Ahmadi (2010), using 1-p IRT, examined gender as a source of DIF in the special English section of IUEE. Their results revealed that IUEE suffers from gender-DIF. Similarly, Ahmadi and Darabi Bazvand (2016), investigating gender-DIF of IUEE employing both 1-p IRT and logistic regression models, found gender to be a source of DIF on the test. Alavi and

Bordbar (2017) also explored gender as a source of DIF in IUEE, employing a Rasch model approach. Based on their findings, the researchers maintained that IUEE is not free from construct-irrelevant variance and its fairness, hence, is jeopardized as the results revealed gender DIF. In contrast, Salehi and Tayebi (2011), examining gender-DIF in the reading comprehension section of an IUEE through employing three-step logistic regression procedure, reported no significant test-item bias toward neither males nor females.

To date, as far as the researchers are concerned, only a few empirical studies have put forward the examinees' field of study as a source of DIF in IUEE. Using the 1-p IRT model, Barati, Ketabi, and Ahmadi (2006) examined performance on the general English section of IUEE among test takers from mathematics, humanities, and sciences backgrounds. The results uncovered 33 DIF items across various sections of the general English subtest. The DIF items in the word order and vocabulary sections favored test takers from sciences or humanities. Half of the DIF items in the cloze test, reading comprehension, language function, and structure sections favored those with a mathematics background, and the rest favored those with sciences and humanities backgrounds. In the same vein, Ravand et al. (2019) examined the general English subtest of master of English programs IUEE to identify possible gender and field of study DIF in the items of this test through employing multiple-indicators multiple-causes structural equation modeling on test performance data of 21,642 applicants. The results of data analysis through Mplus showed reveled 12 items being flagged for DIF in a statistically significant way. However, only five items showed DIF in terms of practical significance.

Considering the paucity of research on the field of study as a source of DIF and test bias in IUEE, the present study was undertaken to probe if the items in the reading section of an IUEE developed for MA applicants of English Majors showed DIF toward the participants. In this respect, the present research endeavored to address two research questions as follow:

Research Question One: Does the reading comprehension section of the general English test in MA University Entrance Exam for English majors show Differential Item Functioning (DIF) in favor of any field of study?

Research Question Two: If so, how much do the findings of DIF detection methods of 1-p logistic and 2-plogistic IRT model converge or differ?

3. Methodology

3.1. Participants

The participants targeted for the present study were chosen from a population of 5000 applicants, including 3927 females and 1073 males, who sat for the MA IUEE held in 2017 for English Majors. The applicants were categorized as belonging to one of the following four groups according to their field of study: Group one, including 1973 applicants from English Translation (ET), Group two, including 1159 applicants from majors other than English, Group three, including 971 applicants from English Language and Literature (ELL), and Group four, including 897 applicants from Teaching English as a Foreign Language (TEFL). For selecting the participants of this study, disproportionate stratified sampling was employed as it provides better results compared to proportionate stratified sampling when one is interested in investigating differences among various strata (Ary, Jacobs, & Razavieh, 1996). Accordingly, from the whole population of 5,000 applicants who sat for the test, an equal number of participants was randomly chosen from the four groups of applicants described above. To this end, the group with the smallest sample size was selected as the basis (N=897). That is, all the participants in the TEFL group were kept. Then the sample size of the other three groups was reduced to 897. The final sample included 3588 participants, selected from the four groups (ELL, ET, TEFL, non-English majors) with 897 applicants in each group.

3.2. Instruments

To collect the required data, the students' responses to one version of IUEE were examined. IUEE for MA in English majors has a total of 240 items with 60 General English questions coming under

four subtests: (a) Structure (10 items), (b) Vocabulary (20 items), (c) Cloze Passage (10 items), and (d) Reading Comprehension (20 items), and 180 Specialized English items coming under three subtests: (a) TEFL (60 items), (b) ELL (60 items), and (c) ET (60 items). The test items are all in multiple-choice format, which are dichotomously scored. Guessing effect is minimized in IUEE as examinees are penalized for their wrong responses. In other words, every three wrong responses will eliminate the correct response. In brief, in this scoring system, negative marking is used to prevent the testees from guessing and boost the test qualities of reliability and validity. In the present research, only the reading comprehension section of the general English subtest containing 20 items was dealt with as our goal was to investigate the potential role of the field of study in the examinees' attempt to answer the reading comprehension questions. It should be mentioned that the General English subtest is the section that all examinees sit for whether their field is TEFL, ELL, ET, or majors other than English. The reliability reported for the reading comprehension section, measured using Cronbach's alpha reliability estimate, was .74.

3.3. Data Analysis Procedure

The data for this research were gained from the National Organization for Educational Testing (NOET), Tehran, Iran, which is responsible for preparing and administering major nation-wide examinations held in the country. Having taken the consent of NOET, the researchers of the present study were provided with the anonymous answer sheets belonging to the applicants of MA IUEE in English majors, held in 2017 by NOET.

The data of the study were assessed by implementing 1-p and 2-p logistic IRT models of DIF detection. As the guessing effect is controlled in the IUEE test, employing 1-p and 2-p IRT models in this study seems appropriate, and there is no need to utilize a 3-p IRT model. The reason why both 1-p and 2-p IRT models were used in this study is the debate that exists over the advantages of the 1-p IRT model in comparison to the 2- and 3-p IRT models. It is argued that in the 1-p IRT model, if item discrimination differs, there is a danger of mis-estimation (McNamara & Roever, 2006). However, the 1-p IRT model is more frequently used in comparison to the 2- and 3-p IRT models since the employment of the 1-p IRT model is more feasible as it requires a smaller sample size in comparison to the 2- and 3-p models (Camilli & Shepard, 1994). Therefore, when a larger sample size is available, the results would be more informative when the 2- or 3-p IRT models are used as they take into account more parameters. Furthermore, the 1-p IRT model is mathematically and technically less challenging than the other two IRT models.

In line with the recommendation put forward by Mizumoto and Plonsky (2015) regarding the more frequent application of R statistical software by quantitative researchers, for analyzing the data of this study, the difR package developed by Magis et al. (2010) was run. In spite of being mostly disregarded in the applied linguistics field, use of R has been gaining rapid momentum in other disciplines (Loewen et al., 2014), and as stated by Muenchen (2014), nowadays R has become the most prevalently utilized analytics software for scholarly papers.

The difR is an R package used for analyzing DIF through nine different methods. According to Magis et al. (2010), using difR package has the following advantages: First, the different methods can be set up with a similar structure and feature and many flexible options for the user. Second, the package handles several DIF detection methods, so they can be compared in one run. Third, the package has been developed for the R software, and it can thus be obtained for free. The package requires some knowledge of the R environment; the help manual provides additional information and references for the interested user. (p. 859)

To be assured of the appropriateness of IRT models, it is imperative to check the underlying assumptions. The first assumption deals with unidimensionality of the test, which assumes that all items within a test must assess a single outstanding trait. However, unidimensionality is not a strict concept as a restricted construct may be produced as a result of strict unidimensionality (McNamara, 1996). Consequently, to reach an acceptable model fit, only an approximation to the unidimensionality assumption which is reasonably good is required (McNamara, 1996). As Reckase (1979) stated, a test can be regarded as unidimensional and be subjected to IRT models for analysis

on condition that the first factor explains approximately 20 percent of the variance besides being many occasions larger than the second factor. To check the unidimensionality of the test, the data were subjected to factor analysis.

The second assumption of the IRT models deals with local independence of test items. In this study, the local independence assumption was partly confirmed due to the results of unidimensionality analysis. The logic behind this conclusion is that "the assumptions of unidimensionality and local independence are related in that items found to be locally dependent will also appear as a separate dimension in a factor analysis" (Reeve, 2003, p. 12). Hence, if the first assumption of the IRT models pertaining to test unidimensionality is satisfied, the local-independence assumption will also be satisfied (Reeve, 2003).

4. Results

Since the 1-p and 2-p IRT models require unidimensionality of the test items, an Exploratory Factor Analysis (EFA) was run to probe the assumption. According to Table 1, the total variance extracted by the first factor was 17.76. As noted by Reckase (1979), a test can be considered as unidimensional and be subjected to IRT models if the first factor accounts for almost 20 percent of the variance besides being larger than the second factor.

Table 1: Total Variance Explained of the Reading Section Items

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.553	**17.765**	17.765	3.553	17.765	17.765	2.128	10.641	10.641
2	1.138	5.691	23.456	1.138	5.691	23.456	1.861	9.306	19.947
3	1.050	5.249	28.705	1.050	5.249	28.705	1.738	8.689	28.636
4	1.035	5.175	33.881	1.035	5.175	33.881	1.049	5.244	33.881
5	.983	4.917	38.797						
6	.971	4.855	43.652						
7	.949	4.744	48.396						
8	.917	4.585	52.981						
9	.900	4.499	57.480						
10	.893	4.464	61.943						
11	.852	4.260	66.203						
12	.839	4.196	70.400						
13	.820	4.102	74.501						
14	.795	3.975	78.477						
15	.773	3.867	82.344						
16	.760	3.799	86.143						
17	.741	3.704	89.847						
18	.718	3.589	93.436						
19	.679	3.393	96.829						
20	.634	3.171	100.000						

Table 2 displays the results of the 1-p IRT model. The results indicated that three items; i.e., items four ($b = 2$, $\chi^2 = 12.48$, $p < .05$), six ($b = 2.56$, $\chi^2 = 9.57$, $p < .05$) and 11 ($b = 1.97$, $\chi^2 = 14.29$, $p < .05$) exercised DIF toward the examinees based on their field of study. As displayed in Item Characteristic Curve (ICC) 1 (Figure 1) and Appendix A, the guessing parameter for all of the items was set at zero. That is to say, a 1-p IRT model assumes that the guessing parameter – lower asymptote – is zero; i.e., the left tail of the curve meets the probability of zero.

Table 2: 1-P Logistic IRT Model of The Reading Comprehension Items

Item	Chi-Square	P	b(difficulty) Parameter	SE(b)
1	5.147	0.161	1.7827	0.0494
2	2.400	0.494	2.3371	0.0566
3	3.569	0.312	1.8943	0.0506
4	12.485	0.006*	2.0055	0.0519
5	4.495	0.213	2.9426	0.0682
6	9.571	0.023*	2.5690	0.0605
7	1.435	0.697	2.3633	0.0570
8	5.140	0.162	2.1544	0.0539
9	3.763	0.288	1.2087	0.0446
10	1.093	0.779	2.5624	0.0604
11	14.294	0.003*	1.9701	0.0515
12	1.836	0.607	2.9910	0.0693
13	1.857	0.603	3.1372	0.0730
14	0.509	0.917	2.5824	0.0607
15	1.274	0.735	2.5134	0.0595
16	6.104	0.107	2.2256	0.0549
17	0.303	0.960	2.2972	0.0560
18	1.992	0.574	1.7868	0.0494
19	1.671	0.644	2.7013	0.0630
20	2.136	0.545	2.4077	0.0577

As indicated in Table 2, the pattern of the curve indicated that item four was a very difficult item for all groups (b = 2). That is to say, a minimum ability of two is needed for having a 50 percent chance to attempt the item correct. According to Figure 1, if a line is drawn from the vertical axis – the likelihood of correct answer – to the curve and then it is continued vertically to cut the horizontal axis – difficulty parameter– the point will be the ability needed to answer the item correctly with a probability of 50 percent. On the upper left end of the curve, the able students missed the item due to difficulty, carelessness, fatigue, or time-out (slipping parameter). The right end of the curve did not touch the probability of 1, another sign that item four was a difficult one.

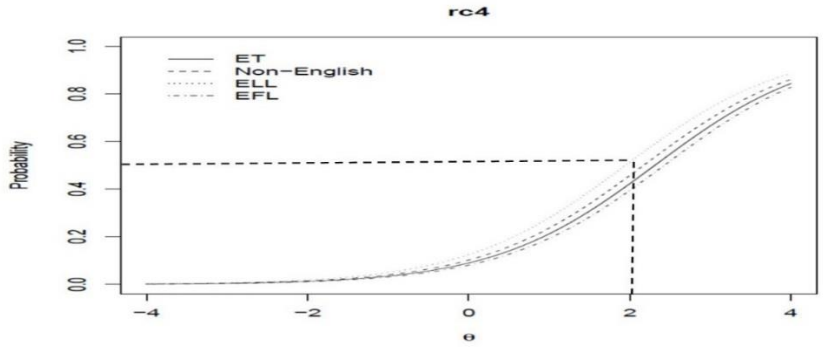


Figure 1: Item Characteristic Curve of Item Four

Based on the comparison of the performance of the groups on this item, it can be claimed that ELL examinees had the best performance on item four, while TEFL students found it more difficult than the other groups. In other words, the able examinees in the latter group missed the item more than the other groups.

As displayed in Table 2 and ICC Curve (Figure 2), the difficulty parameter for item six was 2.56. That is to say, item six was more difficult than item four (b = 2). Item six was the easy one for the TEFL students. ELL and non-English students found it almost equally difficult. It should also be noted that item six had a larger slipping parameter. The right upper ends of the curves were lower than the probability of .80.

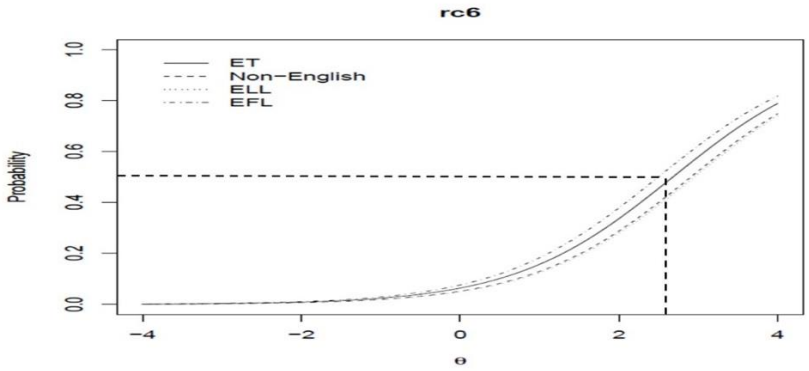


Figure 2: Item Characteristic Curve of Item Six

Compared to item four and six, as can be noticed in Figure 3, item 11 was easier (b = 1.97). English Language and Literature (ELL) students had the best performance on item 11, while the other three groups found it almost equally difficult.

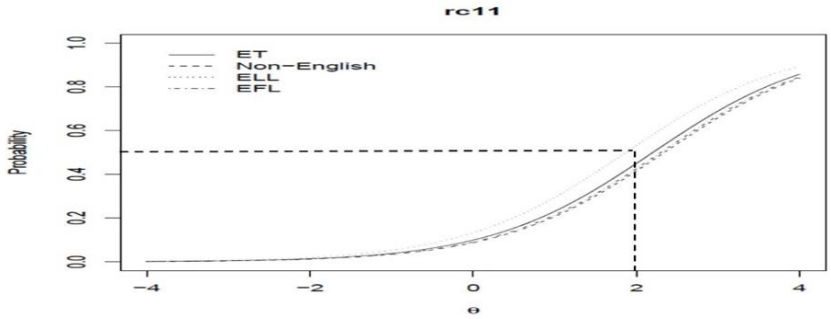


Figure 3: Item Characteristic Curve of Item 11

Table 3 demonstrates the outcomes of the 2-p IRT model. In a 2-pl model, the difficulty (a) and discrimination (b) parameters are computed for each item. The results indicated that, using a 2-p logistic IRT model, all items of the reading section exercised DIF toward examinees based on their field of study. Three of the items with the highest DIF were; items 18 (a = 1.79, b = 1.24, χ^2 = 71.70, p < .05), 16 (a = 1.12, b = 1.99, χ^2 = 69.58, p < .05) and four (a = 1.54, b = 1.49, χ^2 = 48.65, p < .05).

Table 3: Two-Parameter Differential Item Functioning

Item	Chi-Square	P	a(Discrimination) Parameter	b(Difficulty) Parameter	SE(a)	SE(b)	Cov (a, b)
1	26.966	0.000	1.2505	1.5048	0.0761	0.0711	-0.0044
2	21.511	0.002	1.1680	2.0496	0.0829	0.1106	-0.0082
3	15.617	0.016	0.4882	3.3533	0.0570	0.3688	-0.0203
4	48.653	0.000	1.5401	1.4982	0.0922	0.0621	-0.0045
5	24.720	0.000	1.3350	2.3512	0.1057	0.1322	-0.0128
6	14.389	0.026	0.9538	2.6022	0.0807	0.1794	-0.0136
7	33.643	0.000	1.7865	1.6250	0.1125	0.0632	-0.0057

8	38.401	0.000	0.9524	2.1837	0.0719	0.1355	-0.0089
9	47.578	0.000	1.1093	1.1078	0.0644	0.0588	-0.0027
10	23.455	0.001	1.0848	2.3616	0.0849	0.1440	-0.0113
11	30.947	0.000	1.3124	1.6110	0.0815	0.0745	-0.0050
12	16.429	0.012	1.0315	2.8572	0.0950	0.2089	-0.0188
13	20.869	0.002	1.3679	2.4621	0.1140	0.1442	-0.0151
14	16.138	0.013	1.1892	2.2334	0.0894	0.1259	-0.0102
15	38.977	0.000	1.4213	1.9443	0.0974	0.0922	-0.0078
16	69.583	0.000	1.1281	1.9996	0.0791	0.1086	-0.0076
17	34.414	0.000	1.7995	1.5774	0.1117	0.0605	-0.0053
18	71.700	0.000	1.7900	1.2490	0.1004	0.0469	-0.0033
19	35.892	0.000	1.4808	2.0337	0.1053	0.0975	-0.0090
20	43.216	0.000	2.2879	1.4925	0.1465	0.0499	-0.0054

As displayed in the following ICC Curve (Figure 4), the results showed that, compared to the other three groups, ET students found item 18 an easy one. However, ELL students found item 18 a difficult one, as the left tail of their curve was the longest of all. Furthermore, the curve was the steepest one for the ET group. That is to say, item 18 showed the highest discrimination toward this group. The results also indicated that item 18 could have been a misleading one for the non-English students. While the weaker students had a higher chance to answer the item correctly, the able non-English students missed it more than the other groups. Besides, item 18 showed the least discrimination toward the non-English group as the curve was the flattest one.

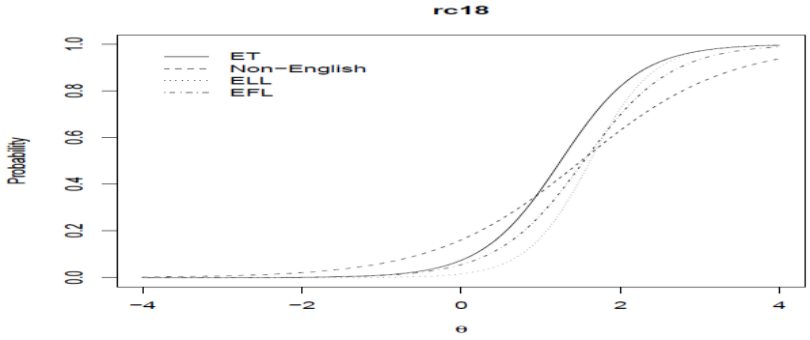


Figure 4: Item Characteristic Curve of Item 18

ICC Curve (Figure 5) displays the performance of the four groups on item 16. On the left side of the curve, item 16 was the easiest for the weak Non-English group, while it was almost equally difficult for the other three groups. On the right side of the curve, item 16 was misleading for the non-English group more than the other groups. Moreover, this item had the highest discrimination toward the ET group as shown through its steep curve. However, it had the lowest discrimination toward the non-English group as it had the flattest curve.

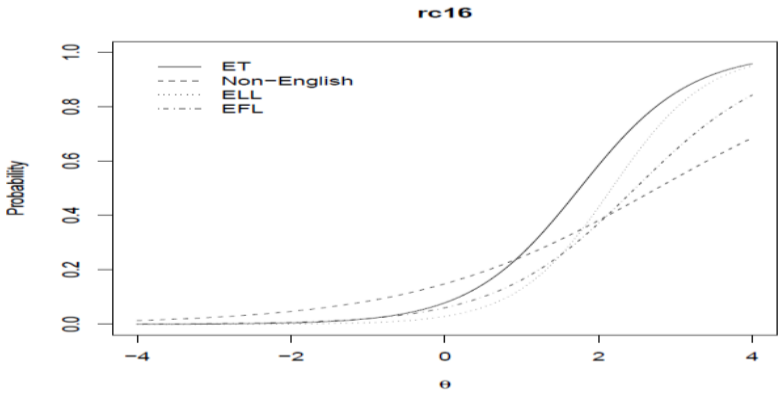


Figure 5: Item Characteristic Curve of Item 16

ICC Curve (Figure 6) displays the performance of the four groups on item four. On the left side of the curves, item four was the easiest for the weak non-English group, while it was almost equally difficult for the TEFL and ELL groups. On the right side of the curve, item four was misleading for the ET group more than the other groups. Item four showed the highest discrimination toward the ELL group as shown through its steep curve. It had the lowest discrimination toward the non-English group that had the flattest curve. Appendix B presents the ICC curves for the other items using a 2-p IRT model.

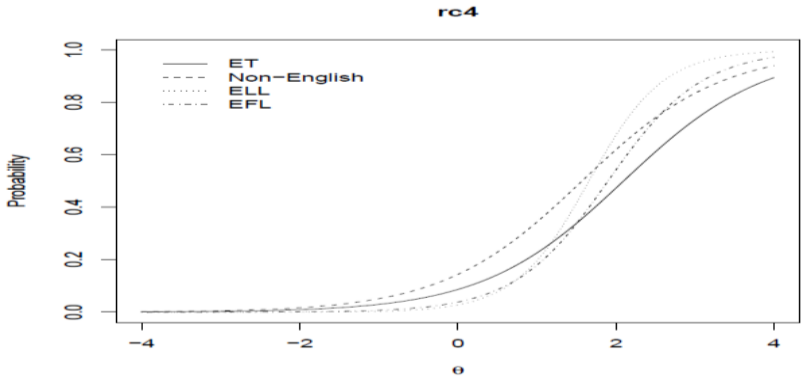


Figure 6: Item Characteristic Curve of Item Four

5. Discussion

The present study attempted to examine potential DIF in the reading comprehension subtest of IUEE considering the test takers' academic background. To this aim, two research questions were raised. With regard to the first research question, it was found that only three items (i.e., items 4, 6, and 11) showed DIF toward examinees based on their field of study. As maintained by McNamara and Roever (2006), "a single differentially functioning item on a multi-item test is not likely to make the entire test unfair. Only when the effect of a number of DIF items accumulates can this lead to a biased test" (pp. 84-85). Accordingly, when analyzed through 1-p IRT model in which only item difficulty parameter was considered (Hambleton et al., 1991), the reading section items did not show much bias as only three out of 20 items were flagged for DIF.

Concerning item four, it was a difficult item for all groups as the difficulty of the item was two. However, the ELL group had the best performance and the TEFL group had the worst performance on this item among the four groups. Regarding item six, this item was even more difficult for all groups compared to item four, as the difficulty of the item was 2.56. In contrast to item four, TEFL group had the best performance on item six. Additionally, ELL and non-English students found it almost equally difficult. As for item 11, this item was easier compared to item four and six as its difficulty was 1.97. Among the four groups, the ELL group had the best performance on the item, while the other three groups found it almost equally difficult. All in all, when examined through the 1-p IRT model, the items were more or less the same for students of various fields of

study as only two items were less difficult for the ELL group and one item was less difficult for the TEFL group, and the remaining 17 items were of approximately the same level of difficulty for all groups.

Concerning the second research question, it was found that all 20 items of the reading comprehension section were found to show DIF. Accordingly, when analyzed through the 2-p IRT model that considers both parameters of item discrimination and item difficulty, all items revealed DIF. Considering McNamara and Roever's (2006) remark, the conclusion made is that the reading comprehension subtest is biased. Among the 20 items, items 4, 16, and 18 showed the highest levels of DIF with regard to examinees' academic background. Concerning item 18, the results indicated that besides being less difficult for the ET group in comparison to the other groups, the item showed the highest discrimination toward the ET group. Furthermore, the item showed the least discrimination toward the non-English group, and it could have been a misleading one for this group. With regard to item 16, it was revealed that the item had the highest discrimination toward ET group, and it had the lowest discrimination toward the non-English group. Furthermore, item 16 was more misleading for the non-English group compared to the other groups. Concerning item four, it was demonstrated that the item showed the highest discrimination toward the ELL group and the least discrimination toward the non-English group.

All in all, the results of 2-p IRT analysis were in line with those of other researchers who maintained that test-takers' academic background can lead to a differential performance in language tests (Barati et al., 2006; Hale, 1988; Pae, 2004; Ravand et al., 2019; Song et al., 2015). Moreover, Camilli and Shepard (1994) found that the 2-p IRT model was more informative and accurate as it could identify more items showing DIF. The justification for this discrepancy between the results of the 1-p and 2-p IRT models can be that the 2-p IRT model scrutinizes items more carefully through considering both parameters of item discrimination and item difficulty. When examined through the 1-p IRT model in which the item discrimination is set at zero, only three items were flagged for DIF due to their differential level of difficulty toward specific groups of test-takers. However, when investigated through the 2-p IRT model, in which items were considered for both their item difficulty and discrimination level, DIF was shown on all items.

The results of the 2-p IRT model also indicated that all items showed the least discrimination toward the non-English group (Appendix B). It should be mentioned that the parameter of item discrimination explains how sharply the item discriminates between examinees with similar ability levels. That is, a high-discriminatory item discriminates very accurately between examinees with very similar ability levels, while a low-discriminatory item indicates distinctions among a wider range of ability levels (McNamara & Roever, 2006). Given the limited exposure to English language in EFL contexts compared to ESL contexts, the English proficiency level of the test takers may be higher than the non-English group as they should pass various courses in English to take a BA degree in English. Thus, high-discriminatory items toward English groups identified possible nuances of difference between them as their ability levels might not be that much dispersed in comparison to the non-English major group who might come under a wider range of language proficiency levels.

6. Conclusion and Implications

The vast majority of DIF studies have considered gender and ethnicity in high-stakes tests (McNamara & Roever, 2006). To uncover the impact of other potential causes of DIF, the current study considered the effect of field of study as a source of DIF on the performance of MA applicants in IUEE of English majors using the 1- parameter and 2- parameter IRT model. The results indicated that when analyzed through the 1-p IRT model, only three items showed DIF; however, when analyzed through the 2-p IRT model, all the 20 items of the reading comprehension section were flagged for DIF. Therefore, it can be safely concluded that first, 2-p IRT model is more accurate than 1-p IRT model as it could identify more DIF items, and second, the reading comprehension subtest of IUEE is biased toward examinees from different fields of study.

The outcomes of this research can be beneficial to stakeholders, test developers, and test administrators, especially those who are in charge of running high stakes tests, as through employing IRT models, they can identify various sources of potential DIF existing in their tests and consequently, attempt to produce test items which are fair to examinees with various academic, gender, ethnic, and national backgrounds. Nevertheless, the results of this research must be cautiously generalized to high stakes tests in general as the data had been provided from the one-time administration of IUEE in 2017. Likewise, these results should be wisely generalized as the assumption of local independence was confirmed through the findings of the unidimensionality analysis. In this study, only IRT models were employed. Future research studies can use other DIF detection techniques and compare their results against those of the present study to come up with more solid findings.

By drawing on the outputs of this research, future researchers are recommended to use those IRT models which consider the highest number of parameters. Provided that their sample size be big enough and their data meet the assumptions behind IRT models, the researchers are suggested to use 2-p and 3-p IRT models as they have the potential to map more items showing DIF and consequently, provide a more accurate picture of the results. Besides, as mentioned by Pae (2004), DIF studies examining the role of the field of study are very scarce. Thus, more studies of this type are recommended to do DIF analysis in various high-stakes tests with the goal of enhancing test fairness. Finally, DIF studies are recommended to focus on the interaction of field of study with other factors such as cultural and language backgrounds, gender, and native language to hopefully broaden our understanding of DIF and its potential causes in high-stakes tests.

References

Ahmadi, A., & Darabi Bazvand, A. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research, 4*(1), 63-82.

Ahmadi, A., & Jalili, T. (2014). A confirmatory study of differential item functioning on EFL reading comprehension. *Applied Research on English Language, 3*(2), 55-68.

Alavi, S. M., & Bordbar, S. (2017). Differential item functioning analysis of high-stakes test in terms of gender: A Rasch model approach. *Malaysian Online Journal of Educational Sciences, 5*(1), 10-24.

Ary, D., Jacobs, L. C., & Razavieh, A. (1996). *Introduction to research in education* (*5th ed.).* Fort Worth: Harcourt Brace College Publishers.

Barati, H., & Ahmadi, A. R. (2010). Gender-based DIF across the subject area: A study of the Iranian national university entrance exam. *The Journal of Teaching Language Skills, 2*(3), 1-26.

Barati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high-stakes tests: The effect of field of study. *Iranian Journal of Applied Linguistics, 9*(2), 27-49.

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

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

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathematics and Social Sciences, 50*(199), 45-59.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *South African Journal of Industrial Psychology, 30*(4), 52-58.

DeMars, C. E. (2010). *Item response theory*. New York, NY: Oxford University Press.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368.

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.

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.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104-109.

Hale, G. (1988). Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing, 5*(1), 49-61.

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hope, D., Adamson, K., McManus, I.C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge-based assessment. *BMC Medical Education, 18*(6), 1-8.

Kane, M., & Bridgeman, B. (2017). Research on validity theory and practice at ETS. In R. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 489-552). Cham, Switzerland: Springer.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59-76.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115-136.

Linn, M. C., De Benedicts, T., Delucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: What does "I don't know" really mean? *Journal of Research in Science Teaching, 24*(3), 267-278.

Loewen, S., Lavolette, E., Spino, L., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly, 48*(2), 360-88.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862.

Martinková, P., Drabinová, A., Liaw, Y., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE-Life Sciences Education, 16*(2), 1-13.

McNamara, T. F. (1996). *Measuring second language performance.* London, UK: Addison Wesley Longman.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Basil Blackwell.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-34.

Mizumoto, A., & Plonsky, L. (2015). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics, 37*(2), 284-291.

Muenchen, R. A. (2014). *R passes SPSS in scholarly use, Stata Growing Rapidly*. Retrieved January 4, 2019 from http://r4stats.com/2014/08/20/r-passes-spss-in-scholarly-use-stata-growing-rapidly/.

Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly, 4*(2), 149-164.

Oliveri, M. E., Lawless, R. R., Robin, F., & Bridgeman, B. (2017). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education, 31*(1), 1-16.

Pae, T. (2004). DIF for students with different academic backgrounds. *Language Testing, 21*(1), 53-73.

Pae, H. (2011). Differential item functioning and unidimensionality in the Pearson Test of English Academic. Retrieved February 14, 2019 from *www.pearsonpte.com/research/Documents/Pae.pdf*

Ravand, H., Firoozi, T., & Rohani, G. (2019). Investigating gender and major DIF in the Iranian National University Entrance Exam using multiple indicators multiple-causes structural equation modelling. *Issues in Language Teaching (ILT), 8*(1), 33-61.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230.

Reeve, B. B. (2003). An introduction to modern measurement theory. Retrieved January 27, 2019 from *http://appliedresearch.cancer.gov/areas/cognitive/immt.pdf*.

Salehi, M., & Tayebi, A. (2011). Differential item functioning (DIF) in terms of gender in the reading comprehension subtest of a high-stakes test. *Iranian Journal of Applied Language Studies, 4*(1), 135-168.

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

Shohamy, E. (2000). Fairness in language testing. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 15-19). Cambridge, UK: Cambridge University Press.

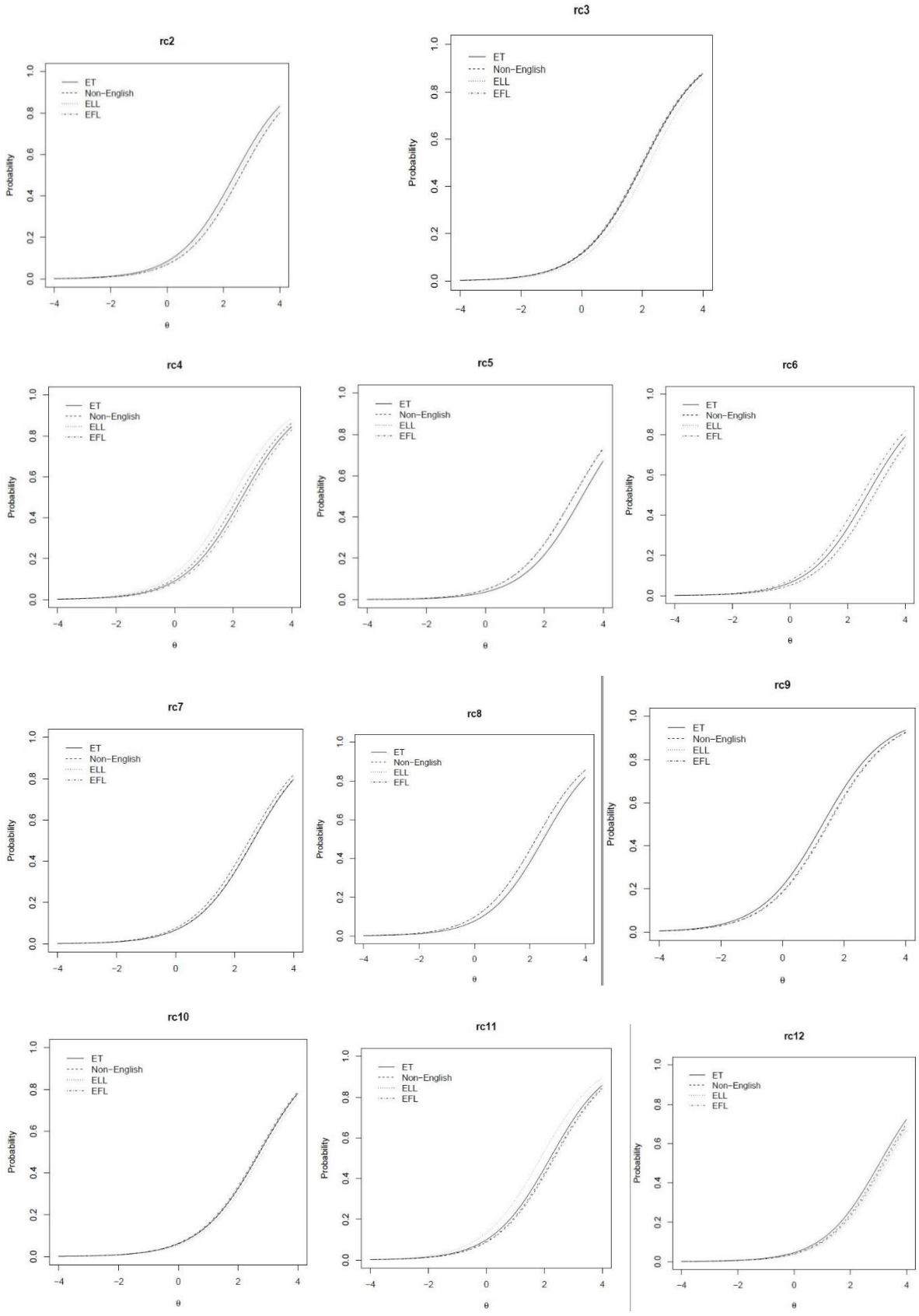
Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Papers in Language Testing and Assessment, 4*(1), 97-124.

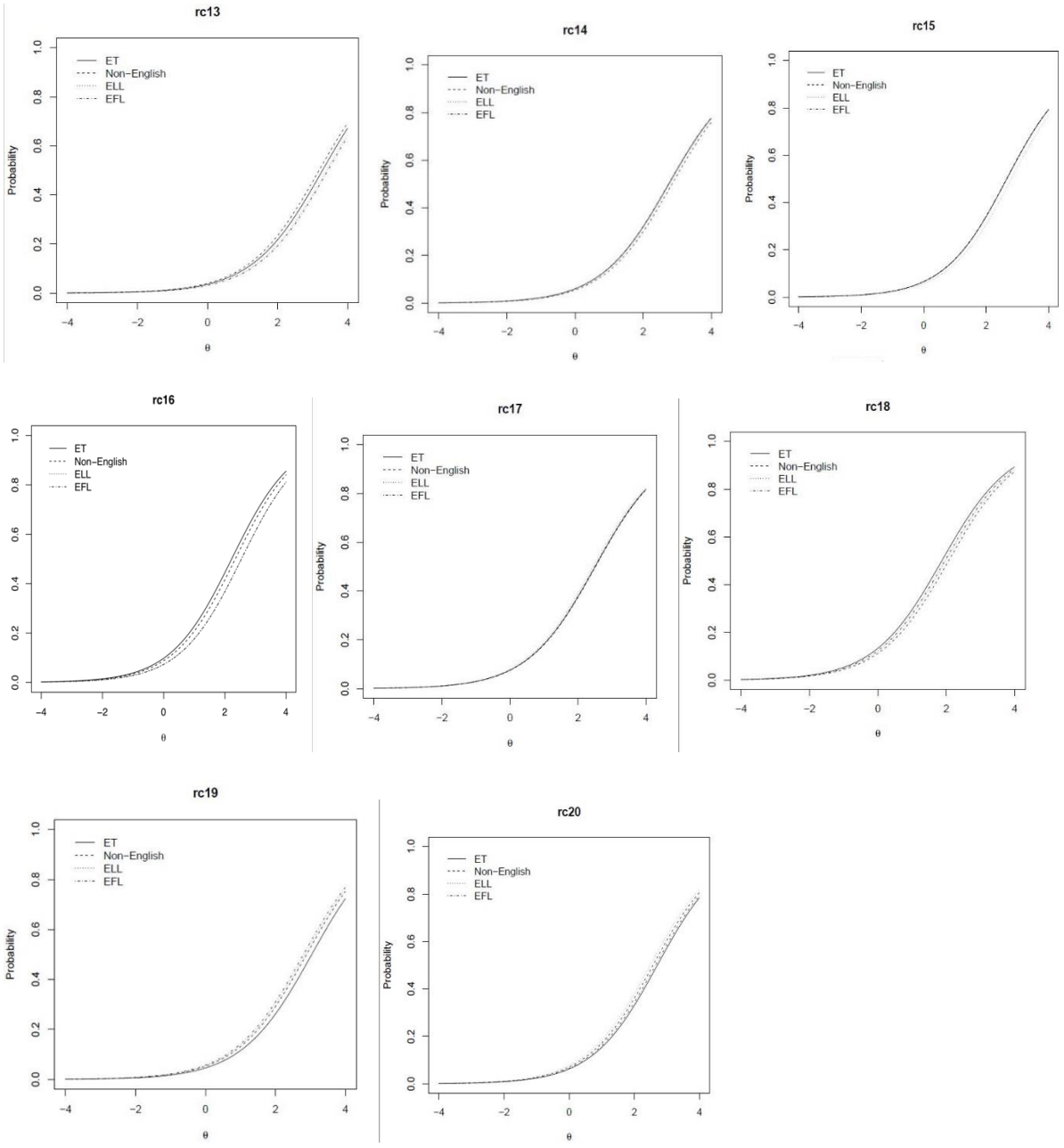
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17(*3), 323-340.

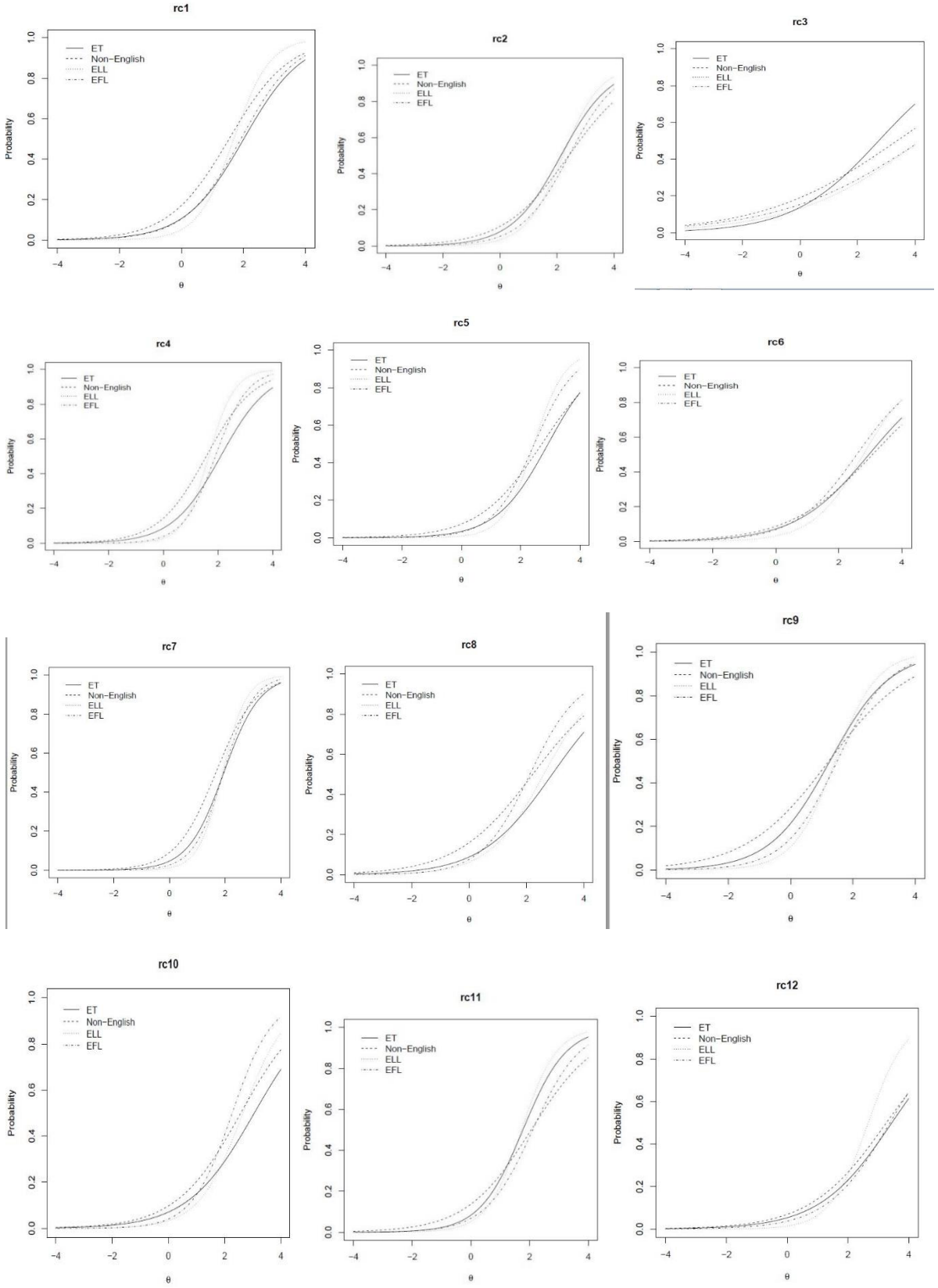
Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Low School Admission Test as an example. *Applied Measurement in Education, 8*(2), 157-187.

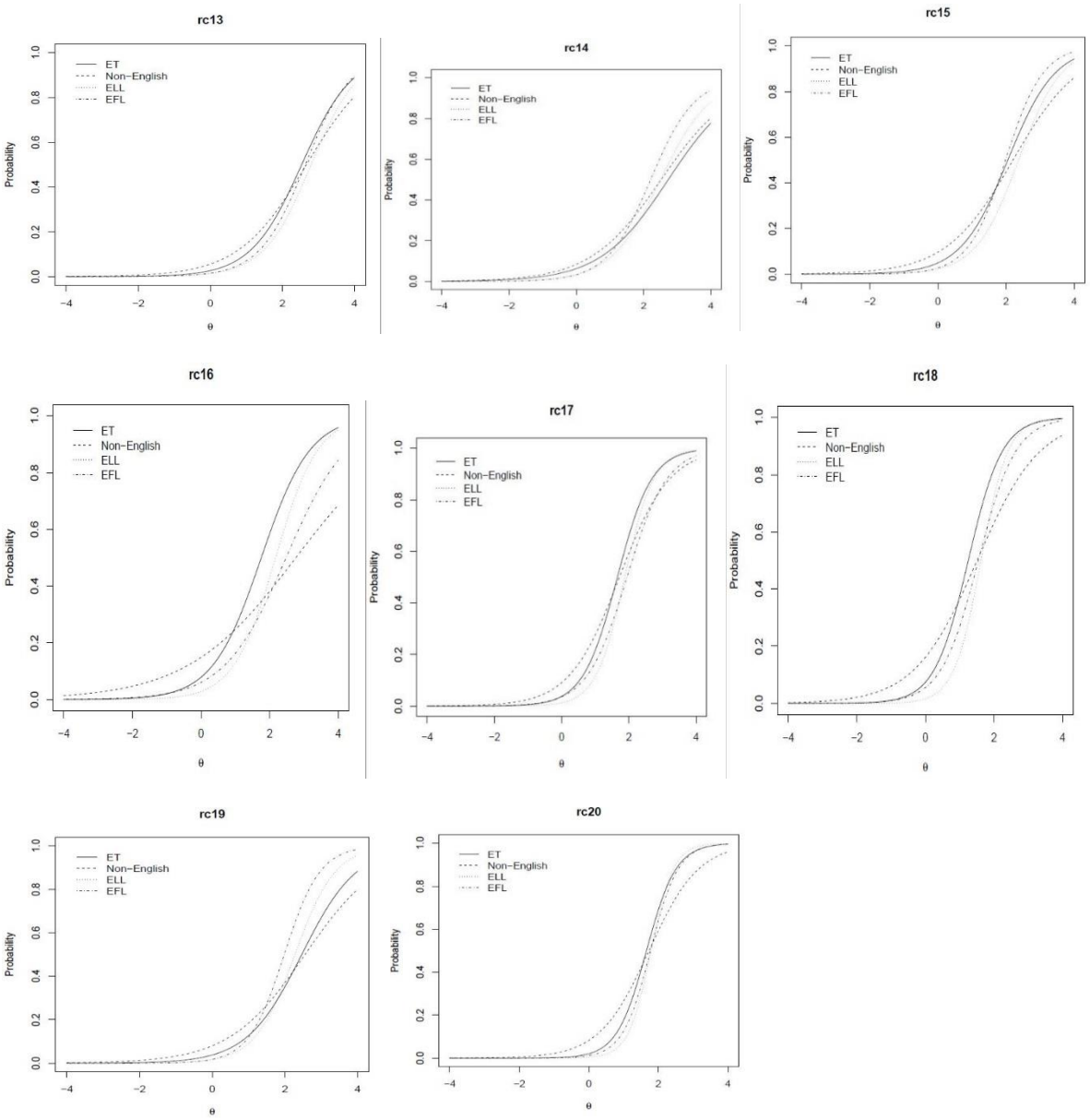
Appendix A: ICC Curves: 1-p IRT model




rc13
rc14
rc15
rc16
rc17
rc18
rc19
rc20
ET
Non−English
ELL
EFL
Probability
θ

Appendix B: ICC Curves: 2-p IRT model




rc13
rc14
rc15
rc16
rc17
rc18
rc19
rc20
ET
Non−English
ELL
EFL
Probability
θ