Journal of English language Teaching and Learning

University of Tabriz

Volume 14, Issue 30, (Fall and Winter 2022)

Native language-based DIF across the subtests:
A Study of the Iranian National University Entrance Exam

Parviz Ajideh

English Language & Literature Department, University of Tabriz, Tabriz, Iran.

parvizaj@gmail.com

Massoud Yaghoubi Notash

English Language & Literature Department, University of Tabriz, Tabriz, Iran.

masoud.yaghoubi@gmail.com

Hamidreza Babaee Bormanaki (Corresponding Author)

English Language & Literature Department, University of Tabriz, Tabriz, Iran.

hreza86b@gmail.com

ARTICLE INFO:

Received date:
2022.05.29
Accepted date:
2022.08.07

Print ISSN: 2251-7995
Online ISSN: 2676-6876

Keywords:

Iranian Undergraduate University Entrance Special English Exam (IUUESEE), Rasch Analysis, Differential Item Functioning, Native language.

Abstract

This paper reports on an investigation of native language-based differential item functioning (DIF) across the subtests of Iranian Undergraduate University Entrance Special English Exam (IUUESEE). Fourteen thousand one hundred seventy two foreign-language test takers (including four groups of Azeri, Persian, Kurdish, and Luri test takers) were chosen for the study. Uniform DIF (UDIF) and Non-uniform DIF (NUDIF) analyses were conducted on data from the four versions of IUUESEE. After establishing the unidimensionality and local independence of the data, DIF findings showed that Luri test takers were more advantaged than other native language groups across the subtests. NUDIF analysis uncovered that almost all subtests functioned in favor of low-ability test takers who haven’t been expected to outperform high-ability test takers. A probable explanation for native language-ability DIF was that Luri and low-ablity test takers were more likely to venture lucky guesses. Thoughtless errors and guessing, test-wiseness, overconfidence, stem length, unappealing distractors, and time were proposed as possible causes of DIF in IUUESEE. It was also found that the reading subtest included the large number of items with significant DIF.

DOI: 10.22034/ELT.2022.51852.2491

Citation: Ajideh, P.; Yaghoubi Notash, M.; Babaee Bormanaki, H. (2022). Native language-based DIF across the subtests: A Study of the Iranian National University Entrance Exam. *Journal of English Language Teaching and Learning*, 14(30), 39-56. Doi: 10.22034/ELT.2022.51852.2491

1. Introduction

Investigation of test fairness is a significant enquiry to decrease or remove bias and discrimination against some groups of test takers, providing them with the equal opportunities to demonstrate their knowledge and skills, and increasing social justice (Gipps & Stobart, 2009; McNamara & Ryan, 2011). In the context of second language proficiency testing, investigation of the fairness of high-stake tests is of paramount importance because they play an important role in test taker' lives. For that reason, the development of the high-stake tests should go through a meticulous process of item analysis in order to validate that all participants with the same level of language abilities have the equal probabilities of correctly answering the items (Camilli and Shephard, 1994).

Differential item functioning (DIF) is a statistical tool for examining test fairness. It investigates the extent to which a test function differently across different groups. DIF is generated when the probability of answering an item correctly is different by groups of participants with the same level of language proficiency (Thissen, Steinberg, & Wainer, 1993). DIF analysis is a fundamental requirement for validity arguments for supporting inferences from test outcomes (American Education Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014).

This study examines the DIF of the Iranian Undergraduate University Entrance Special English Exam (IUUESEE) across its subtests. The IUUESEE was launched in 1999 by the Iranian National Organization of Educational Testing. The test is in multiple choice format including structure, vocabulary, word order, language function, cloze, and reading comprehension sections.

In the context of language testing, statistical DIF analysis including Rasch-based procedures have been the topic of research among researchers who have tried to investigate the fairness of tests in order to disclose statistical bias in test items (see, e.g., Muraki, 1999; Roznowski & Reith, 1999; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000; Zenisky, Hambleton, & Robin, 2003; Zhang, Matthews-Lopez, & Dorans, 2003). Standardized fit statistics and Rasch mean square (MNSQ), have been typically applied to Rasch-based analysis for examining the applicability of the data set to the Rasch model. In this regard, different ranges of acceptable fit indices have been proposed by different researchers (Bond & Fox, 2007; Linacre, 2010) (see the DIF analysis section for more discussion). The current study used Rasch-based DIF procedures to investigate the proportion of native language-based DIF across the subtests of IUUESEE.

2. Literature Review

2.1 DIF analysis

As a Rasch-based study, the current research drew on differential item functioning (DIF) which is a common method in the language testing for examining bias. Presence of DIF indicates an interaction between the test takers' performance and a characteristic (e.g., native language, gender, age, nationality, or race), implying an unfair advantage to the specific group of test takers (see Kunnan, 1990; Zeidner, 1986, 1987). In order to have meaningful effect on the measurement, DIF must be statistically significant ($p < .05$) (Linacre, 2010).

This study is categorized as the “second DIF generation” framework (Zumbo, 2007). In psychological and educational contexts, multiple methods have been developed to examine DIF including the Rasch model, the Mantel–Haenszel procedure, multidimensional item response theory, the Standardization procedure, logistic regression, and etc. In this study, we applied the Rasch model to our DIF analysis.

DIF is categorized as either UDIF or NUDIF (Ferne & Rupp, 2007). We choose this model because it can identify both uniform DIF (UDIF) and Non-Uniform NUDIF (Linacre, 2010). Other methods can only identify UDIF, excluding logistic regression (Swaminathan, 1994). Investigation of NUDIF is of particular importance and it hasn’t often been examined in DIF studies and many of the studies which have not revealed UDIF, have been disclosed to have NUDIF (see Mazor et al., 1994). Negligence in investigation of NUDIF may lead to critical practical consequences (Ferne & Rupp, 2007).

UDIF indicates that “there is no interaction between ability level and group membership” (Prieto Maranon, Barbero Garcia, & San Luis Costas, 1997, p. 559), meaning that when the four native language groups function in different ways on a given test item, their differences continue to be constant across all test takers’ ability levels. This specifies that item characteristic curves (ICCs) of two subgroups form equal slopes but differing intercepts which refers to a constant difference between the two subgroups (e.g., Azeri and Persian), regardless of the ability levels under investigation (Aryadoust. et. al., 2011). On the other hand, NUDIF indicates that performance differences fluctuates across test takers’ ability levels. Specifically, the performance differences among the native language groups will not remain constant among the different ability levels of those groups. This points to the interaction of native language with ability levels which leads to “nonparallel item characteristics curves” (Prieto Maranon et al., 1997, p. 559). This also forms different slopes which leads to the intersection of ICCs (Zumbo, 1999).

In this study, we also investigated unidimensionality which according to Ferne and Rupp (2007), investigates whether the overall test scores contaminated by any extraneous factor, and local independence examining the effect of test takers’ performance on a test item on their performance on another item. (Ferne & Rupp, 2007). This perspective is called multidimensionality-based DIF analysis by Roussos and Stout (1996, 2004) because it relates DIF analysis to dimensionality analysis. This approach relates the underlying causes of significant DIF to the presence of multidimensionality in items (Ackerman, 1992; Shealy & Stout, 1993). The multidimensional DIF approach enable researchers to take account of the secondary dimensions which is not related to the construct dimension (Geranpayeh & Kunnan, 2007). For that reason, dimensionality analysis is a significant precondition for Rasch-based DIF analysis (Ferne & Rupp, 2007, p. 129). Only eight studies of twenty seven DIF studies reviewed by Ferne and Rupp (2007), investigated the evidence of unidimensionality. In the current study, we investigated both unidiminsionality and local independence by applying multidimensional approach to the test data.

This study is exploratory in nature. Initially, we identified the items with significant UDIF and NUDIF across the native languages and the subtests. Then, we tried to generate hypotheses regarding the causes of DIF, explaining the findings through previous studies and the evidences found from the analyses.

2.2 Previous research on the effect of native language on test outcomes

Reviewing the pertinent literature reveals two groups of studies investigating the effect of language background on test takers' performance on standardized tests. The first group includes studies which examined the variance between test takers' performance at the item level (Alderman and Holland, 1981; Chen and Henning, 1985; Sasaki, 1991; Ryan and Bachman, 1992; Kim, 2001; Uiterwijk and Vallen, 2005; Harding, 2011). Chen and Henning (1985) and Sasaki (1991) found that vocabulary subtests in different tests favored test takers with Spanish as a native language. By employing the Rasch model to calibrate item difficulty estimates and plotting them across native speakers of Chinese and Spanish, Chen and Henning (1985) found that DIF items found from the vocabulary subtests favored Spanish test takers. They related the reasons for DIF to cognate words. In a similar vein, Sasaki (1991) examined the UCLA English as a Second Language Placement Examination (ESLPE) and uncovered that vocabulary items with English–Spanish cognates displayed DIF against the Chinese language group, whereas items with idiomatic expressions functioned in favor of the Chinese test takers. They related the instructional background of the Chinese Test takers to DIF results.

Using the Mantel–Haenszel method, Ryan and Bachman (1992) found DIF in TOEFL subtests, with some functioning in favor of the non-Indo-European (NIE) Group and others advantageous to Indo- European (IE) group. More recently, Harding (2011) revealed that Japanese L1 listeners were favored on a small number of items on the listening subtest of University Test of English as a Second Language (UTESL) containing the Japanese-accented speaker, whereas Mandarin Chinese L1 listeners were noticeably favored on large number of items on the test containing a Mandarin Chinese L1 speaker.

The second group of these studies examined the constructs of different language tests across different language groups (Swinton and Powers, 1980; Oltman *et al*., 1988; Hale *et al*., 1989; Kunnan, 1994; Ginther and Stevens, 1998; Brown, 1999; Ackerman *et al*., 2000). As a matter of fact, this line of research investigated the extent to which a test measures the same factor structures among groups of test takers with different native languages (Kim, 2001). In one of the early studies, Swinton and Powers (1980) found different constructs comparing non-Indo-European (NIE) and Indo-European (IE) test takers on the Test of English as a Foreign Language (TOEFL). In a similar vein, Kunnan (1994) also identified different constructs affecting test performance across IE and NIE groups by comparing two different structural models which provided different model-fits. On the other hand, some researchers *(Oltman et al., 1988; Hale et al., 1989; Brown, 1999; Ackerman et al., 2000)* identified same factor structures of tests across different native language groups. For instance, Hale et al.'s (1989) found a similar factor structure including listening and nonlistening in the TOEFL across four different language families, specifically Semitic, Sino-Tibetan, Altaic and Indo-European languages. Ackerman *et al*. (2000) investigated dimensionality comparing Korean, Arabic, and French test takers on the TOEFL listening comprehension section and revealed one dimension across the three groups. By applying generalizability theory, Brown (1999) discovered the similar magnitudes of variance across 10 different language groups. It appears that there is an inconsistency between these two groups of test takers who investigated the performance of different language groups at the test level.

In spite of examining the effect of native language on test performance from different perspectives, these studies had some shortcomings including the unbalanced small sample size and short tests, generalizability of findings (being conducted in western countries), applying arbitrary criterion for identifying DIF and lack or scarcity of studies examining DIF across the ability levels. These limitations left gap in DIF research and the current study has tried to help fill this gap by employing large sample size and the large number of items, conducting the study in Iranian context to generalize the findings to the eastern countries, and analyzing both UDIF and NUDIF.

2.3 Previous research on IUUESEE

In recent years, researchers have studied DIF of IUUESEE in terms of gender and field of study: Barati and Ahmadi (2010) reported that females were favored on three subtests of the test including grammar, language function, and cloze, while males were favored on the vocabulary and word order subtests. Furthermore, the reading comprehension was found to function in favor of both males and females equally. They have also found that reading and vocabulary subtests included the largest number of UDIF items (16 and 17 items) and word order the fewest (3 items). Brati et. Al. (2006) found the similar statistics with reading section including 14 and word order section including 3 UDIF items with reference to test takers' fields of study.

However, the test has not yet been exposed to native language-based UDIF and NUDIF analyses across the subtests. With this considerations in mind, the objective of the preset study was to investigate the effect of native language on item functioning and the proportion of UDIF and NUDIF across the IUUESEE subtests. This together with the crucial role that the IUUESEE plays in the educational lives of Iranian students served as a motive for the current study to examine the fairness of IUUESEE by comparing four native language groups of Azeri, Persian, Kurdi, and Luri across six subtests by means of Rasch analysis. In order to address this purpose, this study specifically addresses the following research questions:

1. Does the Rasch analysis provide evidence of unidimensionality and local independence in IUUESEE?

2. Does the test data fit to the Rasch model?

3. Does the IUUESEE include UDIF items comparing Azeri, Persian, Kurdi, and Luri native language groups? If so, to what extent does the test function differentially across the four groups and to what extent do the different subtests of IUUESEE include the proportions of UDIF instances and items?

4. Does the IUUESEE include NUDIF items comparing high-ability and low-ability levels of the four native language groups? If so, to what extent does the test function differentially across the ability levels and to what extent do different subtests of IUUESEE include the proportions of NUDIF instances and cases?

5. What are the possible causes of UDIF and NUDIF in IUUESEE?

3. Method

3.1 Test materials

The research instrument was the IUUESEE (the Iranian Undergraduate University Entrance Special English Exam). The IUUESEE is an English proficiency test developed by the National Organization of Educational Testing in Iran. Each version of this test consists of six subtests , a total of 70 items in length which includes structure (10 items), vocabulary (15 items), word order (5 items), language function (10 items), cloze test (15 items), and reading comprehension (15 items).Versions 2016, 2017, 2018, and 2019 were chosen for the current study.

3.2 Participants

Overall, a total sample of 14172 participants were selected for the present study. All participants were learning English as a foreign language, and represented a range of four native language (L1) backgrounds: Azeri, Persian, Kurdish, and Luri (Table 1).

Table 1. *Number of participants by first language (L1)*

	Native language				
Test Versions	**Azari**	**Persian**	**Kurdi**	**Luri**	**Total**
2016	1213	1349	493	374	3429
2017	1076	1306	495	364	3241
2018	1329	1570	606	472	3977
2019	1252	1377	485	399	3114
Total	4870	5602	2079	1069	14172

3.3. Data collection procedure

The National Organization of Educational Testing provided us with the anonymous answer sheets for the test takers of IUUESEE 2016, IUUESEE 2017, IUUESEE 2018, and IUUESEE 2019. This organization design, organize and administer national examinations in Iran such as the university entrance exam for high school graduates, university entrance exam for MA candidates, and etc.

3.4. Data analysis

Data analysis was conducted in three phases: 1. Analysis of descriptive statistics, item difficulty and person ability measures, fit to the Rasch model, and reliability. 2. Examination of unidimensionality and degree of local independence of data 3. Analysis of UNDIF and NUDIF. Analysis of descriptive statistics was conducted by means of Excel 2013 for Windows. WINSTEPS computer program, version 5.1 (Linacre, 2021) was used to conduct the Rasch-based analyses including fit, reliability, item and person measures, point-measure correlation, unidimensionality and local independence, and DIF.

3.4.1 The Rasch model

The Rasch model is a data analysis procedure which creates multi-item interval scales by assigning test takers and test items on a continuum on which the position of test takers and items tallies to their ability and difficulty estimates, respectively (Aryadoust, 2012). Therefore, item difficulty and person ability are two important components of this model. Item difficulty is estimated with reference to the number of test takers who respond to the item correctly and person ability is estimated with regard to the proportion of items that are answered correctly.

In this way, the Rasch model estimates the probability of a test taker's response to an item correctly by variance between test taker ability and item difficulty. In this regards, if a particular test taker's ability is higher than the difficulty of a particular test item, he (she) will probably answer the item correctly (Wright & Stone, 1988).

Fit analysis is one of the important parts of the Rasch model. It examines the extent to which the test data fit to the rasch model. Infit MNSQ and Outfit MNSQ are reported in this study based on the suggestion by M. Linacre in the Winsteps manual (2012). According to Linacre and Wright (1994), Infit MNSQ is an inlier-sensitive information-weighted index which is affected by the deviations from expected patterns in test items near average difficulty. Outfit MNSQ is an outlier-sensitive index influenced by the deviations from expected patterns in test items of low or high difficulty (Linacre, 2002). We expect the MNSQ value of 1.0, therefore, for instance, a value of 1.2 includes 20 percent noise more than the amount expected by the model increasing the standard error of measurement (Smith, 1996; Wright & Linacre, 1994). Linacre (2010) suggest an acceptable range of fit indices between 0.5 and 1.5. Bond and Fox (2007) divided items into two groups of underfitting and overfitting indices. MNSQ indices are greater than 1.4 in underfitting items and less than 0.6 in overfitting items.

Rasch analysis is also used for examining the reliability of the test for both items and persons. Reliability indices range from 0 to 1 in Rasch analysis. Low reliability points to the contamination of variability in measures driven by a high standard error of measurement (SEM). As an another index for reliability, separation is also estimated which points to the ratio of test items' or test takers' standard deviation to their root mean square standard error (Linacre, 2010). Separation ranges from zero to infinity.

We also estimated point-measure correlation for all items of the four test versions as part of our Rasch analysis. By examining the consistency between observed scores and the construct, point-measure correlations reveal the degree of the uniformity between them.

3.4.2 Test of Dimensionality and local independence

Dimensionality analysis is used to examine whether a test item measure the same latent trait and if this condition is met, the test is proved to be unidimensional. Unidimensionality is an indispensable requirement for Rasch-based DIF analysis (see Linacre, 2010, for more explanation). Principal component analysis of linearized Rasch residuals (PCAR) is used to examine the Unidimensionality in this study. Residuals are resulted from the variance between the expectations of the Rasch model and the observed data (Linacre, 1998a; Wright, 1996). PCAR searches for unexpected part of the data which is not consistent with Rasch measures and this unexpectedness results from the same pattern shared by the group of items (Linacre, 2012). These items may also share a substantive attribute in common which is called a "secondary dimension" (Linacre, 2012, p. 553). The presence of secondary dimensions in the data is an evidence of multidimensionality of the test under investigation.

Statistical independence in data happens when the value of one datum has no effect on the value of another (Wright, 1996). In this study, local independence is examined via Pearson correlation analysis of linearized Rasch residuals. Residuals, which are the variance between the observed difficulty measure of items and the values projected by the Rasch model, disclose "how much locally easier or harder that item was than expected" (Wright, 1994b, p. 510).

3.4.3 Differential Item Functioning Investigation

In this study WINSTEPS software, version 5.1 (Linacre, 2021) is used to investigate both UDIF and NUDIF effect size. According to Linacre (2010) effect size points to the ratio of the variance in local item difficulty between subgroups to the standard deviation of the reference group. Effect size is "insignificant" when it is below 0; when it approximates 0.4, it becomes "slight to moderate"; and when it is higher than 0.6, it is considered as "moderate to large" (Linacre, 2010, p. 487). The local difficulty measures are also compared by native language through Welch t test according to the p value of .05 (p < .05) based on Linacre (2010). As mentioned before, we made use of the Rasch model to examine native language-based UDIF and NUDIF. The reason for examining NUDIF is to test for its replication in order to know whether the items still persist to show DIF in the same way for each subgroup and if it does, the DIF signifies the sign of real DIF and If not, it is likely to be just an indication of sampling issue (Du, 1995). Therefore, after we found items with significant UDIF, we divided each native language group into two subgroups of high-level and low-level test takers to see whether DIF continues to exist. The recurrence of DIF in this phase indicates systematic DIF (Du, 1995).

4. Analysis and Results

4.1 Individual descriptive and Rasch analyses of IUUESEE 2016-2019

The results of the individual Rasch analyses for IUUESEE 2016-2019 are demonstrated in table 2. These preliminary statistics were estimated to scrutinize the technical quality of the four versions of IUUESEE. This preliminary investigation determines the extent to which the test items fit to the Rasch model, the reliability of the test, and etc. As shown in table 2, each IUUESEE administration had an average mean score.

Mean raw scores for IUUESEE 2016-2019 were 0.49, 0.45, 0.45, and 0.45 respectively meaning that they functioned similarly in examining test takers' level of language proficiency. The standard deviation of each IUUESEE administration was also 0.52 on average, which indicates that the scores were spreading out moderately.

The items comprising the four versions of the IUUESEE fit to the Rasch model to some extent, which, in this regard, Infit and Outfit MNSQ values for all items ranged from 0.06 to1.48 and from 0.59 to 3.28 respectively. In these ranges, a number of items with high and low MNSQ values didn't accord to our fit criteria. Closer look into these items revealed that they were either very easy or very difficult. Guessing or thoughtless errors by some test takers were probably the reasons for theses unexpected scores.

Item and person reliability ranged from 0.49 to 1, which indicates a moderate degree of replicability. The person reliability of IUUESEE 2018 is lower than that of IUUESEE 2016, 2017 and 2019. Item separation which is an estimate of the separation or spread of the items along the measured variable, ranged from 10.24 to 15.79. This shows that the items can be separated into 10 to 15 statistically distinct strata of difficulty. Person separation ranged from 0.98 to 1.29, meaning that the participants can be divided into 1 or two statistically different strata of performance according to their scores on the IUUESEE.

Table 2 *Descriptive and Rasch statistics for IUUESEE*

Measures	IUUESEE 2016	IUUESEE 2017	IUUESEE 2018	IUUESEE 2019
Mean (raw score)	0.49	0.45	0.45	0.45
SD (raw score)	0.46	0.46	0.45	0.44
Mean Rasch person measure	-0.35	-0.51	-0.45	-0.44
Mean Rasch item measure	-0.0024	-0.00014	-0.0029	-0.0075
Skewness	-0.30	0.24	0.18	0.23
Kurtosis	-3.11	-1.26	-1.13	-0.66
Maximum infit MNSQ	1.48	1.33	1.3	1.42
Minimum Infit MNSQ	0.78	0.82	0.85	0.06
Maximum outfit MNSQ	3.28	1.61	1.96	2.81
Minimum outfit MNSQ	0.59	0.74	0.67	0.73
PT-Measures	0.41	0.38	0.38	0.346
Item reliability	0.99	0.99	0.99	1
Item separation	12.38	10.24	13.27	15.79
Person reliability	0.63	0.53	0.49	0.56
Person separation	1.29	1.07	0.98	1.26

Note. N = 14172; Number of items = 70 ; PT-Measures = point measure correlation

4.2 Results of uidimensionality and local independence

Table 3 presents the results of unidimensionality analysis conducted by WINSTEPS software. PCAR found that all variances are very close to the Rasch model prediction proving that the estimation of the Rasch difficulty measures was successful (Linacre, 2010). Furthermore, the first contrast in the residuals explains only 2.5% of the variance in the whole data. These findings support the assumption of unidimensionality in IUUESEE.

Table 3 *Variance explained by Rasch*

	variance explained by Rasch dimension		
IUUESEE versions	Observed	Eigenvalue	Expected
2016	28.1%	27.4	27.7 %
2017	23.5%	21.4	23.5%
2018	29.6%	29.4	29.6%
2019	29.7%	29.5	29.7%

Analysis of Pearson correlations significantly supported the assumption of local independence. Correlations above 0.70 indicate local dependence in the data (Linacre, 2010). All correlations in the four IUUESEE versions ranges from - 0.13 to 0.29 supporting the local independence of all items.

4.3 Differential items functioning

We searched for the occurrence and recurrence of DIF across the subtests. IUUESEE includes six subtests: Structure, Vocabulary, Word order, Language function, Cloze, and Reading. We have traced and searched for the number of items favoring each native language group and subgroup in a specific section, the number of times a specific item functioned in favor of a native language group and a subgroup, and the number of UDIF and NUDIF items occurred in each subtest in separate test versions. This section is conducted in two parts. In the first part, we searched for the occurrence and recurrence of DIF across the subtests in items with significant UDIF and in the second part, we conducted this for items with significant NUDIF. Our findings are demonstrated in the following tables.

4.3.1 UNDIF across the subtests

Table 4 presents the number of items which functioned in favor of each native language group in separate test versions across the subtests. For instance, only one item in structure section of IUUESEE 2016 functioned in favor of Azeri test takers.

Table 4 *Native language UNDIF in IUUESEE: Number of items favoring each group in separate test versions*

Subtests	**Test Versions**																	**Total**
	2016				**2017**				**2018**				**2019**					
	A	P	K	L	A	P	K	L	A	P	K	L	A	P	K	L		
Structure	1	2		1	1	1			1	1	2	1	1				2	14
Vocabulary		2	2	2		1	2		4					5	3	3	6	30
Word order		1		1			1		1			2				1		7
Language function	1	2		1	3	2				2				4	2	4	2	23
Cloze		2	1	1	1		1				2		3	2	3	2	3	21
Reading	1	4		4	1	1	2		3	2	2	4	4	2	2	3	2	37
Total	3	13	3	10	6	5	6		9	5	6	7	8	13	10	13	15	132

We have demonstrated the number of items which functioned in favor of each group in all test versions in table 5. As the table shows, six items in the vocabulary section of all test versions favored Persian test takers. The table 5 also shows sum of the items functioning in favor of each group. For example, overall, 42 items of 280 items of all four test versions favored Luri test takers.

Table 5 *Native language UNDIF in IUUESEE: Number of items favoring each group in all test versions*

Test batteries							
Native language	Structure	Vocabulary	Word order	Language function	Cloze	Reading	**Total**
Azeri	3	5	0	10	3	6	27
Persian	5	6	1	6	7	9	34
Kurdish	1	7	4	4	4	9	29
Luri	5	12	2	3	7	13	42
Total	14	30	7	23	21	37	132

Table 6 shows the number of items with significant UDIF in each test section in separate test versions, and sum of the UDIF items in each section and version. It is noted that we only counted one of the DIF instances of items which simultaneously favored more than one native language group. Reading section and test version 2019 included the large number of UNDIF items than the other sections and versions.

Table 6 *Native language UNDIF in IUUESEE: Number of items occurred in each subtest in separate test versions*

Test batteries							
Test versions	Structure	Vocabulary	Word order	Language function	Cloze	Reading	**Total**
2016	4	5	2	3	3	8	25
2017	3	7	2	3	2	5	22
2018	3	0	2	2	4	8	19
2019	2	10	1	6	8	7	34
Total	12	22	7	14	17	28	100

4.3.2 NUDIF across the subtests

We have demonstrated our findings of NUDIF across different subtests in tables 7 and 8. We searched for the number of items with significant NUDIF in each subgroup occurring in each subtest. We only listed one of the NUDIF instances of a NUDIF item for each subgroup. For example, item 35 of test version 2018 function four times in favor of high-ability Azeri test takers compared with four native language subgroups. In cases like these, we counted only one of NUDIF cases of item 35 and didn’t include the other cases. Therefore, table 29 presents the number of NUDIF items favoring each subgroup along with their occurrence and recurrence in each section. For instance, 15 items of the structure section of all test versions favored the high-ability Azeri test takers, whereas 24 items of this section favored low-ability Azeri test takers. Overall, 100 items favored high-ability Azeri test takers based on NUDIF analysis.

Table 7 *Native language NUDIF in IUUESEE: Number of items favoring each subgroup in all test versions*

Test batteries							
Native language subclasses	Structure	Vocabulary	Word order	Language function	Cloze	Reading	**Total**
Azeri 1	15	20	8	17	17	23	100
Azeri 2	24	20	10	14	24	21	113
Persian 1	15	21	6	15	19	32	108
Persian 2	20	21	9	18	21	21	110
Kurdish 1	10	18	7	16	21	24	96
Kurdish 2	16	21	10	12	12	16	87
Luri 1	12	21	7	15	18	24	97
Luri 2	7	6	0	4	8	8	33
Total	119	148	57	111	140	169	744

In table 8, we showed the number of NUDIF items in each subtest considering all test versions. In this table, we only included one of the cases of a NUDIF item in each subtest. The table shows that structure section of exam 2016 included 9 items with significant NUDIF.

Table 8 *Native language NUDIF in IUUESEE: Number of items occurred in each subtest in separate test versions*

Test batteries							
Test versions	Structure	Vocabulary	Word order	Language function	Cloze	Reading	**Total**
2016	9	13	5	8	10	15	60
2017	9	14	5	9	10	10	57
2018	8	10	4	9	14	13	58
2019	9	14	5	8	11	13	60
Total	35	51	19	34	45	51	235

5. Discussion

The present study investigated the effect of native language on DIF in the Iranian Undergraduate University Entrance Special English Exam (IUUESEE). Specifically, this study set out to investigate the preconditions for Rasch-based DIF analysis (i.e., fit, point measure correlation, unidimensionality, and local independence) in four versions of the IUUESEE, the proportion of native language-based UDIF and NUDIF across the IUUESEE subtests and their possible causes, and the extent to which the strict Rasch fit criteria can reveal the presence of DIF. Analysis of descriptive statistics, item difficulty measures, fit, PCAR, Pearson correlations of residuals fulfilled the requirements of Rasch analysis.

Our analyses found strong support for the item reliability and separation of the test, however the person reliability was under question based on Linacre (2012). This finding indicates that IUUESEE resulted in lower ability ranges and has not probably separated high performers from low performers appropriately in terms of the construct which the test was supposed to measure (Linacre 2012). This finding is confirmed by Fit and DIF results. The item reliability and separation coefficients of IUUESEE were high meaning that the test tested the wide range of difficulty, pointing to the largeness of our sample which is able to precisely detect the items on the latent variable (Linacre 2012).

PCAR analysis revealed that IUUESEE is unidimensional and locally independent to some extent. It was also found that Wright and Linacre's (1994) fit criterion (ranged from 0.8 to 1.2) was more advantageous than other criteria such as Bond and Fox (2007) in investigation of test takers' response patterns. In this regard, fit values of many of the items were 1 or close to 1 pointing to the absence of erratic response patterns in the data, although a number of items with high and low MNSQ values didn't accord to our fit criteria. These misfitting items either overfit or underfit the model to some extent, generating unexpected variance among test takers which is probably owing to carelessness or guessing (Wright & Linacre, 1994).

Considering all versions with regard to UDIF, reading section had the largest number of UDIF items (28 items) and word order section had the fewest number of UDIF items (7items). Vocabulary, cloze, language functions, and structure sections included 22, 17, 14, and 12 UDIF items respectively. This finding is in line with our NUDIF results which revealed that reading and vocabulary sections both included the largest number of NUDIF items (51items) and word order section consisted of the fewest (19 items). Our finding is consistent with findings from previous empirical studies on IUUESEE (Barati & Ahmadi, 2010; Barati, Ketabi, & Ahmadi, 2006).

The finding that the small number of the DIF items belong to word order section may be due to the overall fewer number of word order items in all versions of the IUUESEE, however; there might be two reasons why reading section generally included the largest number of UDIF and NUDIF items across native language groups and subgroups in a similar way. The first reason points to an important issue that besides being a power test, IUUESEE is a speed test requiring students to answer the test items in a short period of time which makes majority of test takers not to finish answering the items of the reading subtest which is always the last subtest in the test. This issue may lead to the fewer differences in the test takers' performances in reading subtest from different language groups and subgroups. Therefore, test takers' performances might be influenced by the speediness of the test, not by the language knowledge and skills.

The second reason is that reading has always been the most important skill taught and trained more than any other skill in Iranian secondary schools. Consequently, Iranian English textbooks has been reading-based for many years. Therefore, test takers have been familiar with this skill from the initial stages of English language learning at schools. As a result, it appears that sufficient training and preparation in this skill has diminished the impact of native language among some native language groups and subgroups as far as DIF is concerned.

The study has also provided evidence that item format alone might not explain DIF adequately. Since the IUUESEE is only constructed in multiple choice format, the focus of this study was on MC item format. In this regard, in general, we can conclude that item format alone cannot lead to DIF, instead; we also need to take account of subject area of test items. (or the interaction of item format and subject area). This study revealed that, with regard to UDIF and UUDIF results, different native language groups and subgroups were favored on the different subtests of IUUESEE versions. For instance, in reading subtest, only 8 items favored low-ability Luri test takers compared with 32 items which functioned in favor of high-ability Persian test takers.

Identification of the reasons for observed DIF is a challenging task (Camili & Shepard, 1994; Gierl, 2005), mainly in exploratory DIF research lacking a priori hypothesis (Jang & Roussos, 2009). The results of analysis and reviewing items helped us propose some reasons for the presence of UDIF and NUDIF in IUUESEE. The reason for low-ability and Luri test takers' successful performance compared to their counterparts can probably be related to their successful lucky guesses. Outfit MNSQ patterns of some items confirm this hypothesis. Some items with high item difficulty which were answered correctly by low-ability test takers had outfit MNSQ indices lower than 0.8 and greater than 1.2 meaning that they functioned contrary to our expectation. The reason that high-ability test takers missed some misfitting items is perhaps due to overconfidence, carelessness, and thoughtless errors. This supposition is also confirmed by results of fit analysis indicating that high-ability participants missed easiest misfitting items that their outfit MNSQ values misfit because of sensitivity to outliers.

Test-wisness also can be an assumption underlying the successful performances of low-ability test takers. Test takers who afforded to participate in special classes to practice test taking strategies seemed to have an opportunity to answer items correctly than those who didn't. This highlights the role of test takers' socio-economic status.

6. Conclusion and Future Research

On the basis of the results of UDIF and NUDIF analyses, the current study has provided evidence on the interaction of native language and item functioning. This interaction became more obvious by examining different ability subgroups of each native language group via NUDIF analysis. In this regard, From 280 items in four test versions, 24 comparisons for UNDIF detection, and 7840 comparisons for NUDIF analysis, UDIF and NUDIF analyses respectively revealed 100 and 235 items with significant DIF at the established threshold *p* value of 0.05 suggested by Linacre (2010a). It was found that Luri test takers were favored more on the sections of the Special English Test than other native language groups. Overall, from 6 subtests, 4 subtest functioned more in favor of Luri group than any other native language group which generally included 42 items from 132 items with significant UDIF. NUDIF analyses found that except for reading subtest and low-level Luri test takers, all low-level test takers outperformed high-level test takers across the subtests. Since IUUESEE is in multiple choice format, it is likely that low-ability test takers were encouraged to venture lucky guesses. Other possible causes of DIF might be long stems, unappealing distractors of the large number of items, and less time available to respond to all items. This can be confirmed by the results of fit analyses which disclosed erratic response patterns in the test data. Test-wiseness which is about the test taking strategies relates to test takers' socioeconomic statues seems to be another factor contributing to DIF results.

This study has also found that reading subtest included the largest number of DIF items and word order the smallest. Topic familiarity can relatively justify the DIF in the reading and cloze subtests. It seems they require test takers to rely on their schema to make top-down comprehension processing for understanding the text.

Our study can be classified as Zumbo's (2007) second generation of DIF. To investigate the native language DIF in high-stake tests based on Zumbo's (2007) third generation of DIF, we can examine socio-cultural and contextual factors affecting different native language group's performances. This line of investigation added to a qualitative research which examines the content of the individual items can reveal important information about the interaction between items and native language DIF.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from multidimensional perspective. *Journal of Educational Measurement*, 29(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Ackerman, T.A., Simpson, M.A., & de la Torre, J. (2000). A comparison of the dimensionality of TOEFL response data from different first language groups. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Alderman, D., & Holland, P. (1981). Item performance across native language groups on the TOEFL. *TOEFL Research Report Series*, 9, 1-106. Princeton, NJ: Educational Testing Service.

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*.Washington, DC: AERA Publications.

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. DOI: 10.1080/15434303.2011.628632

Aryadoust, V. (2012). Differential Item Functioning in While-Listening Performance Tests: The Case of the International English Language Testing System (IELTS) Listening Module. International Journal of Listening, *26*(1), 40–60. DOI: 10.1080/10904018.2012.639649.

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*.

Brati, 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.

Brown, J.D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing*, 16 (2), 217–38. https://doi.org/10.1177%2F026553229901600205.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. London, UK: Erlbaum.

Camilli, G., & Shepard, L. (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), 155–63. https://doi.org/10.1177%2F026553228500200204.

Du, Y. (1995). When to adjust for differential item functioning. *Rasch Measurement Transactions*, 9(1), 414.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges and recommendations. *Language Assessment Quarterly*, *4(*2), 113-148. https://doi.org/10.1080/15434300701375923

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. https://doi.org/10.1080/15434300701375758.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, *24*(1), 3–14. https://doi.org/10.1111/j.1745-3992.2005.00002.x.

Ginther, A., & Stevens, J. (1998). Language background and ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In A.J. Kunnan (Ed.), *Validation in language assessment* (pp. 169–94). Mahwah, NJ: Lawrence Erlbaum.

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in 21st century: Connecting theory and practice* (pp. 105-118). Netherlands: Springer Science+Business Media.

Hale, G.A., Rock, D.A., & Jirele, T. (1989). *Confirmatory factor analysis of the Test of English as a Foreign Language*. TOEFL Research Report, 32, 89-42. Princeton, NJ: Educational Testing Service.

Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing,* 29(2) 163–180. https://doi.org/10.1177%2F0265532211421161.

Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing*, *9*(3), 238–259. https://doi.org/10.1080/15305050903107022

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*(1), 89–114. https://doi.org/10.1177%2F026553220101800104.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*(4), 741–746. https://doi.org/10.2307/3587128.

Kunnan, A.J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validation. *Language Testing, 11*(3), 225–52. https://doi.org/10.1177%2F026553229401100301.

Linacre, J. M. (1998a). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266–283.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2010). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2012). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

Linacre, J.M. (2021). Winsteps® Rasch measurement computer program (Version 5.1). Winsteps.com.

Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8*, 350.

McNamara, T., & Ryan, K. (2011). Fairness versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*, 161-78.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel–Haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284–291. https://doi.org/10.1177%2F0013164494054002003.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement*, *36*(3), 217–232. https://doi.org/10.1111/j.1745-3984.1999.tb00555.x.

Oltman, P.K., Stricker, L.J. & Barrows, T. (1988). Native language, English proficiency, and the structure of the Test of English as a Foreign Language for several language groups. *TOEFL Research Report.* 27, 88-26. Princeton, NJ: Educational Testing Service.

Prieto Maranon, P., Barbero Garcia, M. I., & San Luis Costas, C. (1997). Identification of nonuniform differential item functioning: a comparison of Mantel–Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement*, *57*(4), 559–569. https://doi.org/10.1177%2F0013164497057004002.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355–371. https://doi.org/10.1177%2F014662169602000404.

Roussos, L. A., & Stout, W. F. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107–116). Thousand Oaks, CA: Sage.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*(2), 248–269. https://doi.org/10.1177%2F00131649921969839.

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*(1), 12–29. https://doi.org/10.1177%2F026553229200900103.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing* 8 (2), 95–111. https://doi.org/10.1177%2F026553229100800201.

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

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*(3), 516–517.

Swaminathan, H. (1994). Differential item functioning: A discussion. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 63–86). Ottawa, Ontario, Canada: University of Ottawa.

Swinton, S.S., & Powers, D.E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups. *TOEFL Research Report*, 6, 80-32. Princeton, NJ: Educational Testing Service.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, *17*(3), 323–340. https://doi.org/10.1177%2F026553220001700303.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In Holland, P. W. and Wainer, H. W., (Eds.), *Differential item functioning*. (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second-generation immigrants in Dutch tests. *Language Testing, 22*(2), 211–234. https://doi.org/10.1191%2F0265532205lt301oa.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, *9*, 472.

Wright, B. D. (1994b). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, *10*(3), 509–511.

Wright, B. D., & Stone, M. H. (1988). *Identification of item bias using Rasch measurement*. (Research Memorandum No. 55). Chicago, IL: MESA Press.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing*, *3*(1), 80–98. https://doi.org/10.1177%2F026553228600300104.

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*(1), 51–64. https://doi.org/10.1177%2F0013164402239316.

Zhang, Y., Matthews-Lopez, J., & Dorans, N. (2003). *Using DIF dissection to assess effects of item deletion due to DIF on the performance of SAT I: Reasoning sub-populations*. Educational testing Service.

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, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from the British Colombia University Web site: http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf

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. https://doi.org/10.1080/15434300701375832.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing*, *3*(1), 80–98.

Zeidner, M. (1987). A comparison of ethnic, sex and age bias in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing*, *4*(1), 55–71. https://doi.org/10.1177%2F026553228700400106.