Bormanaki and Ajideh
Language Testing in Asia (2022) 12:29
https://doi.org/10.1186/s40468-022-00185-2

Language Testing in Asia

RESEARCH Open Access

Item performance across native language groups on the Iranian National University Entrance English Exam: a nationwide study

Hamidreza Babaee Bormanaki* and Parviz Ajideh

*Correspondence: hreza86b@gmail.com

Department of English, Faculty of Persian Literature and Foreign languages, University of Tabriz, Tabriz, Islamic Republic of Iran

Abstract

This paper reports on an investigation of differential item functioning (DIF) in the Iranian Undergraduate University Entrance Special English Exam (IUUESEE) across four native language groups including the Azeri, the Persian, the Kurdish, and the Luri test takers via Rasch analysis. A total sample of 14,172 participants was selected for the study. After establishing the unidimensionality and local independence of the data, the authors employed two methods to test for the DIF: (a) a *t*-test uniform DIF analysis, which showed that the Luri test-takers were more favored than other native language groups, and (b) nonuniform DIF analysis, which revealed that majority of nonuniform DIF instances functioned in favor of the low-ability Azeri, the low-ability Persian, the high-ability Kurdish, and the high-ability Luri test-takers. A possible explanation for native language-ability DIF was that the Luri and low-ability test-takers were more likely to venture lucky guesses. We also referred to socioeconomic status (e.g., test-wiseness), guessing, overconfidence, thoughtless errors, stem length, time, L1, and unappealing distractors as possible sources of DIF in IUUESEE.

Keywords: Differential item functioning, Iranian Undergraduate University Entrance Special English Exam (IUUESEE), Native language, Rasch analysis

Introduction

In the context of L2 proficiency testing, high-stake tests play an important role in test-takers' lives. This highlights the importance of test fairness, an attempt to rule out or decrease bias against some groups of test-takers providing them with equal opportunities for demonstrating their knowledge and skills, which increases social justice (Gipps & Stobart, 2009; McNamara & Ryan, 2011). Therefore, the development of the high stake tests needs to undergo a rigorous process of item analysis to ensure that all test-takers with the same underlying level of language proficiency have the same probabilities of correctly answering the items (Camilli & Shepard, 1994).

This study evaluates the Iranian Undergraduate University Entrance Special English Exam (IUUESEE) through DIF analysis which is a powerful tool to investigate the statistical bias in test items. The IUUESEE test was established in 1999 by the National Organization of Educational Testing in multiple-choice format and includes structure,

vocabulary, word order, language function, cloze test, and reading comprehension subtests.

The IUUESEE is a high-stake and norm-referenced test which is administered annually for participants whose aim is to be admitted into Iranian foreign language undergraduate programs. Based on their rank in test outcomes, participants can select a university for their education. Because of the paramount importance of IUUESEE which may result in social and personal consequences for the participants, this research will provide new insights into the psychometrics aspects of the test; specifically on the DIF, it may display by individual items. Through this research, the stakeholders, specifically the test designers, will realize the probable effect of test-takers' L1 on test outcomes in different parts of the IUUESEE which can shed light on the construct-irrelevant variance among test-takers and promote the construct validity of the test by giving an opportunity to test designers to revise the subtests and items which may unfairly function in favor of a group (or groups).

The DIF analysis is a statistical technique to estimate the extent to which participants with different aspects but the same level of ability has different probabilities of responding to test items correctly (Cohen & Bolt, 2005; Oliveri et al., 2014; Zumbo, 2007). It shows that some factors apart from the test construct influence the performance of one group but not the other (Timukova & Drackert, 2019). In other words, the DIF is a result of unequal probabilities of correctly answering an item by two groups of test-takers, who are otherwise matched in ability on a construct (Ferne & Rupp, 2007). Therefore, the examination of the DIF is an indispensable step in the validation of educational and psychological tests (Camilli & Shepard, 1994). It provides researchers with a series of techniques to uncover construct-irrelevant factors that are likely to discriminate unfairly against a specific group of test-takers and hence threaten the validity of test outcomes (Pae, 2004).

In the context of test fairness, language testing researchers have used statistical DIF analysis, mainly the RASCH-based procedures to disclose a statistical bias in test items (see, e.g., Aryadoust, 2012; Aryadoust & Zhang, 2016; Belzak, 2019; Timukova & Drackert, 2019; Trace, 2019; Vanbuel & Deygers, 2021; Xuelian & Aryadoust, 2020; Zenisky et al., 2003; Zhang et al., 2003). In this regard, standardized fit statistics and the Rasch mean square (MNSQ) were usually used in Rasch-based investigations to examine the applicability of the data set to the model (for further explanation, see the DIF analysis section). This study applied the Rasch-based DIF analysis to examine native language-based DIF in the Iranian Undergraduate University Entrance Special English Exam.

Literature review

Previous studies on the effect of native language on test-takers' performance

In the context of language assessment, DIF based on language background has been of particular interest to researchers. Two lines of inquiry emerge in the pertinent literature. The first includes studies that examined the structures of a test across different language groups (Ackerman et al., 2000; Brown, 1999; Ginther & Stevens, 1998; Hale et al., 1989; Kunnan, 1994; Li & Suen, 2012; Oltman et al., 1988; Swinton & Powers, 1980). These studies examined whether a test measured the same constructs for several language groups (Kim, 2001). Most directly, Swinton and Powers (1980) identified different

constructs across non-Indo-European (NIE) and Indo-European (IE) language groups on the Test of English as a Foreign Language (TOEFL). On the other hand, Ackerman et al. (2000) examined the dimensionality among three different language groups Korean, Arabic, and French in the TOEFL Listening Comprehension section and identified one single dimension across all three groups.

The second inquiry was comprised of studies that explored the differences in test-takers' performances at the item level (Alderman & Holland, 1981; Chen & Henning, 1985; Harding, 2011; Kim, 2001; Oliveri et al., 2018; Ryan & Bachman, 1992; Sasaki, 1991; Shin, 2021; Shin et al., 2021; Uiterwijk & Vallen, 2005; Xuelian & Aryadoust, 2020). For example, in two early studies into the effect of native language on test performance at the item level, Chen and Henning (1985) and Sasaki (1991) reported that the vocabulary subsection in different tests functioned in favor of the Spanish language groups. Chen and Henning (1985) found that DIF items identified from the vocabulary subsection functioned in favor of the Spanish group rather than the Chinese group. In another study, Oliveri et al. (2018) discovered more DIF items functioning in favor of non-American citizens living in America over American citizens in the verbal reasoning part of the GRE. Recently, Xuelian and Aryadoust (2020) investigated the mother tongue differential item functioning in the Pearson Test of English (PTE) Academic Reading test across Indo-European (IE) and Non-Indo-European (NIE) language families. They found no statistically significant uniform differential item functioning (UDIF) at $p>0.05$; however, they revealed three non-uniform differential item functioning (NUDIF) items out of 10 items across the language families.

Examining the DIF based on the native language would lead to a significant validation inquiry for language test designers in various test situations, especially high-stakes tests (Geranpayeh & Kunnan, 2007). However, a review of these studies revealed several limitations. The majority of native language-based DIF investigations have been conducted in European and American settings (Pae, 2004). Therefore, the generalizability of the findings would be questioned due to the lack of the DIF studies in other settings such as Asian contexts. The current study was carried out in an Asian context, Iran—on four successive versions of the IUUESEE to help fill this gap. These studies detected DIF items with the arbitrary criterion. For instance, In Chen and Henning's DIF analysis, if the confidence interval had been determined narrower than 95%, more instances of significant DIF might have been identified. The unbalanced small sample size and short tests were also problematic. The present study is a nationwide investigation that comprises a large sample size (14,000 test-takers) and a large number of items (70 items). Furthermore, the presence or lack of DIF across the ability levels was not taken into consideration in most of the previous studies. In other words, the procedures employed in those studies did not examine non-uniform DIF (see the "DIF analysis" section). Several studies that have not identified UDIF have been revealed to have NUDIF bias in their test items (see Mazor et al., 1994). In the current study, we used Rasch analysis for identifying both uniform and nonuniform DIF for the dichotomous response items.

Previous research has investigated the DIF of IUUESEE in terms of gender (Barati & Ahmadi, 2010) and field of study (Brati et al., 2006). However, the test has not yet been subjected to native language-based UDIF and NUDIF analysis. Without such analysis, the stakeholders, specifically test developers and test users, are left to suppose that the

test is fair and does not function in favor of any native language group. Therefore, the objective of the present study is to investigate the interaction between item functioning and native language. To address this aim, the study addresses the following research questions:

1. Does the test data support the assumptions of unidimensionality and local independence, as requirements of Rasch-based DIF analysis?
2. Does the IUUESEE contain UDIF items across the Azeri, the Persian, the Kurdi, and the Luri native language groups? If so, to what extent does the test function differentially across the four groups?
3. Does the IUUESEE contain NUDIF items across ability levels of the four native language groups? If so, to what extent does the test function differentially across the ability levels?
4. Can more stringent Rasch fit criteria indicate the presence of DIF?
5. What are the probable factors that caused DIF in items in the test?

Method

Participants

The participants of this study were randomly selected from high school graduates who sat for the IUUESEE in 2016, 2017, 2018, and 2019. Generally, the participants of the IUUESEE are divided into two groups: the first group includes test-takers who take the IUUESEE with an exam of their high school field of study which includes one of the math, science, and literature and humanity fields. The second group includes those who only take the IUUESEE. In other words, this exam is their main exam for entering into undergraduate university programs. The dataset we used in this study contained participants from both groups. The participants of our study were selected from four provinces of Iran according to the four native languages under investigation. Overall, a total sample of 14,172 participants was selected for the current study. Table 1 presents the specific information about the participants.

Materials

Iranian Undergraduate University Entrance Special English Exam

This study evaluates the Iranian Undergraduate University Entrance Special English Exam (IUUESEE). IUUESEE is one of the five exams (math, science, literature and

Table 1 Number of participants by province and first language

	Provinces and native languages				
	East Azerbaijan	**Fars**	**Kurdistan**	**Lorestan**	
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

humanity, special English, and art) of The Iranian National University Exam called the Konkur examination. Konkur is borrowed and changed from the French term "Concours," referring to the process of sourcing, evaluating, and selecting participants for different objectives (Alavi et al., 2021). IUUESEE is a large-scale high-stake standardized test of the English language which was administered in 2002 for the first time throughout the country (Razmjo, 2006). After 20 years, the structure of the exam remained almost intact. The test has contained six subtests which have generally included 70 MC (multiple choice) items: structure (10 items), vocabulary (15 items), word order (5 items), language function (10 items), cloze test (15 items), and reading comprehension (15 items).

The items of the structure section are designed in the form of incomplete sentences which are supposed to be completed by an option from four alternatives. The questions of this section measures test-takers' understanding of a specific grammatical rule or mixture of rules. Vocabulary items are designed in the form of incomplete sentences. The test-takers are supposed to choose the best choice for the completion of the sentence meaning. The word order section includes test items asking test-takers to choose the option which does not include any grammatical mistake relating to the stem of the item. Items in the language functions section ask the test-takers to complete the conversations with the best choice. The cloze section includes a passage comprised of 15 blanks requiring test-takers to select the option which completes the passage. The last section, reading comprehension, includes three reading comprehension texts whose length varies between 350 and 500 words covering a wide range of topics such as academic, scientific, and social issues (Alavi et al., 2021). Each text includes 5 multiple-choice items that check test-takers' understandings of the content of the text.

The time to answer 70 items is 105 min. All items are dichotomous. The exam applies correction for guessing in a way that three incorrect answers would eliminate one of the correct responses (Alavi et al., 2021). The test content is not distributed throughout different subtests equally. For example, structure and grammar contained 27.15%, vocabulary included 34.28%, and reading comprehension comprised 38.57 of the content (Razmjo, 2006).

Data collection procedures

The National Organization of Educational Testing provided the data for the study. This organization is responsible for designing, organizing, and administering national exams such as the university entrance exam for high school graduates and the university entrance exam for MA candidates. The organization provided us with the anonymous answer sheets for the test-takers of the special English exam in 2016, 2017, 2018, and 2019.

Data analysis

In advance of performing DIF analysis, we undertook two main analyses of the test data: 1. an analysis of descriptive statistics, item difficulty measures, fit to the Rasch model, and reliability, and 2. an examination of dimensionality and degree of local independence of dataset.

Descriptive statistics

We estimated descriptive statistics including mean, standard deviation, skewness, and kurtosis coefficients, using Excel 2013 for Windows.

The Rasch model

We conducted the rest of the analysis based on the Rasch model through WINSTEPS, Version 5.1 (Linacre, 2021). In the Rasch model for dichotomous items, there are two core statistical concepts including item difficulty and person ability. The difficulty measure of an item is estimated by taking into account the number of participants who answered the item correctly, regardless of their ability levels and the participant's ability measure is estimated by considering the number of items she (he) answered correctly, regardless of the difficulty level of the items (Linacre, 2012).

Fit

The fit analysis investigates the extent to which the data match the rash model. We reported Infit MNSQ and Outfit MNSQ for items. Considering fit results, Bond and Fox (2007) divided items into two groups: underfitting items and overfitting items. In underfitting items, MNSQ indices are greater than 1.4, and in overfitting items, MNSQ indices are less than 0.6. On the other hand, Wright and Linacre (1994) proposed a rather tough fit criterion which ranged from 0.8 to 1.2. In this study, we preferred Wright and Linacre's criterion because it is stringent and also it can be adjusted to dichotomous data appropriately (Smith, 1996).

Reliability and separation

We used Rasch model to examine the reliability of the test. In the Rasch model, reliability is estimated for both persons and items and ranges from 0 to 1. We also used separation as another index for reliability, referring to the ratio of test items' or test-takers' standard deviation to their root mean square standard error (Linacre, 2010), which varies from zero to infinity.

Point-measure correlation

In this study, point-measure correlations were estimated for all test items. These correlations represent the proportion of the consistency between observed scores and the latent trait (Linacre, 2012). We also estimated the relationships between persons and items on an item-person map or Wright map which represents both person ability and item difficulty along a single line calibrated in log-odd units (logits) (Linacre, 2012).

Unidimensionality and local independence

We estimated unidimensionality through the principal component analysis of linearized Rasch residuals (PCAR). The difference between the expectations of the Rasch model and the observed data leads to residuals (Linacre, 1998; Wright, 1996a). Fit statistics were also used to test for unidimensionality. Test items exhibiting irregular fit indices were purported to include incorrect difficulty measures and were supposed to be affected by a factor not expected by the test designer. The assumption underlying local independence is that response to an item should not affect response to another item in

a test. We tested for local independence using Pearson correlation analysis of linearized Rasch residuals.

DIF analysis

Our study can be classified as the "second DIF generation" framework proposed by Zumbo (2007). The second generation is marked by widespread approval of the term *DIF* rather than *item bias*. In the testing context, multiple methods have been developed for identifying DIF (e.g., Rasch model, the Mantel–Haenszel procedure, logistic regression, etc.). We adopted the Rasch model which has been frequently used in DIF studies. The Rasch model has an important advantage over other methods. It can identify both uniform DIF (UDIF) and non-uniform DIF (NUDIF) (Linacre, 2010). Except for logistic regression (Swaminathan, 1994), other methods can identify only UDIF.

The presence of uniform DIF indicates that an item consistently functions in favor of a particular group of test-takers across all ability levels, and the presence of non-uniform DIF shows that the performance of test-takers varies across the levels of ability (Xuelian & Aryadoust, 2020). In other words, UDIF occurs when "there is no interaction between ability level and group membership" (Prieto Maranon et al., 1997, p. 559). On the other hand, NUDIF is evidence of interaction between ability level and group membership (Golia, 2016). Examination of NUDIF is of paramount importance which is ignored in DIF studies and most of the studies which have not identified UDIF have been found to display NUDIF (see Mazor et al., 1994). Negligence in the investigation of NUDIF may lead to critical practical consequences (Ferne & Rupp, 2007). Therefore, selecting a method of DIF analysis that can uncover both UDIF and NUDIF is of significant importance.

The Rasch model also has the advantage of being able to examine unidimensionality and local independence which, according to Ferne and Rupp (2007), they function as requirements for Rasch-based DIF analysis. Unidimensionality investigates the contamination of overall test scores by any extraneous dimension, and local independence examines whether test-takers' performance on a test item is affected by their performance on another item or not (Ferne & Rupp, 2007). Roussos and Stout (1996, 2004) refer to this perspective as a multidimensionality-based DIF analysis method that integrates dimensionality analysis with DIF analysis. In this approach, underlying causes of significant DIF are related to the presence of multidimensionality in items (Ackerman, 1992; Shealy & Stout, 1993). As Roussos and Stout (2004) stated "such items measure at least one secondary dimension in addition to the primary dimension that the item is intended to measure" (p. 108). This multidimensional paradigm of DIF provides researchers with opportunities to take account of these secondary dimensions (Geranpayeh & Kunnan, 2007). Therefore, dimensionality analysis is a significant requirement for Rasch-based DIF analysis (Ferne & Rupp, 2007, p. 129). In the previous DIF studies, only eight of twenty-seven studies examined unidimensionality (Ferne & Rupp, 2007). However, the current study investigated the unidimensionality and local independence in the test items of IUUESEE to see whether they satisfy the preconditions of DIF analysis or not.

Despite the lack of a comprehensive and solid framework for DIF analysis (see Zumbo, 2007), the majority of the researchers have taken two approaches in their DIF studies over the past decades: (1) Confirmatory approach, in which, at first, hypotheses are

generated through the analysis of test items and then they are tested via DIF analysis (e.g., Gierl, 2005). (2) Exploratory approach, in which at first, researchers explore the items with significant DIF and then they try to generate hypotheses about the causes of DIF and explain the findings through previous studies and the evidence from the results or they try to conduct a posteriori content analysis of the items exhibiting DIF (e.g., Lin & Wu, 2003). A review of 27 studies of DIF analysis revealed that a lot of studies applied exploratory analysis (Ferne & Rupp, 2007). The current study is exploratory. At first, we explored the items with significant DIF and then tried to put forward suppositions regarding the causes of DIF and explain the findings through previous studies and the evidence that were found by analyzing the data.

Results

As a requirement for native language-based DIF analysis of the data, testing for the unidimensionality and local independence in the IUUESEE was the preliminary objective of this study. Therefore, after testing for these statistics, we investigated the presence of UDIF and NUDIF in items that met the stringent Rasch fit criteria proposed by Linacre (2010). What we found is discussed as follows.

Fit of the data to the latent trait model

The results of the descriptive statistics of the test data, as well as the Rasch measurement findings which comprise fit indices, difficulty measures in logits, and point-measure correlations, are in the Appendix. In IUUESEE 2016, item 16 ($M = 0.83$, total score = 1618) and item 21($M = 0.83$, total score = 1649) have the highest mean score, and item 65 ($M = 0.09$, total score = 42) the lowest, indicating that item 16 and item 21 were answered correctly and item 65 incorrectly by majority of test-takers. Item 16 and item 21 were the easiest and item 65 the most difficult. As Table 18 (see the Appendix) shows, in IUUESEE 2017, item 37 ($M = 0.82$, total score = 1249) has the highest mean score, and item 56 ($M = 0.14$, total score = 160) the lowest, which reveals that item 37 was answered correctly and item 56 incorrectly by many of test-takers. In this case, item 37 was the easiest item and item 56 was the most difficult. In IUUESEE in 2018, item 11 ($M = 0.91$, total score = 3200) has the highest mean score, and item 23 ($M = 0.10$, total score = 183) the lowest, which indicates that item 11 was answered correctly and item 23 incorrectly by a lot of test-takers. Therefore, item 11 was the easiest item and item 23 was the most difficult item. In IUUESEE in 2019, item 39 ($M = 0.88$, total score = 2106) has the highest mean score, and item 38 ($M = 0.04$, total score = 108) the lowest, indicating that item 39 was answered correctly and item 38 incorrectly by the majority of the test-takers. Item 39 was the easiest and Item 38 was the most difficult. In all test versions, skewness and kurtosis coefficients fall between −2 and +2 in all items except for item 65 in IUUESEE 2016, item 56 in IUUESEE 2017, items 11 and 23 in IUUESEE 2018, and items 21, 33, 38, 39, and 68 in IUUESEE 2019 which points to univariate normality.

The Infit MNSQ and Outfit MNSQ columns (see the Appendix) demonstrate the test items' infit and outfit MNSQ indices. Fit statistics of items 23, 27, 28, 61, 62, 64, 65, and 66 in IUUESEE 2016; items 24, 25, and 65 in IUUESEE 2017; items 23, 27, 34, 38, 39, 50, 52, 61, and 65 in IUUESEE 2018; and items 5, 25, 38, 51, 65, 68, and 70 in IUUESEE 2019 fall out of the range from 0.6 to 1.4 recommended by Bond and Fox

(2007). In addition to these items, fit statistics of items 5, 7, 9, 12, 15, 20, 22, 38, 48, 55, 56, 57, and 67 in IUUESEE 2016; items 1, 7, 14, 16, 18, 22, 23, 26, 31, 33, 39, 48, 53, 57, 60, and 64 in IUUESEE 2017; items 11, 12, 17, 31, 41, 44, and 68 in IUUESEE 2018; and items 2, 3, 12, 18, 19, 21, 22, 24, 30, 32, 39, 42, 47, 54, 55, 61, and 64 in IUUESEE 2019 fall out of the range from 0.8 to 1.2 recommended by Wright and Linacre (1994). PT-Measures (see the Appendix) which demonstrate point-measure correlations for test items shows that all correlations are positive, except item 65 in test version 2016 and items 25, 38, and 68 in version 2019. These results indicate that there is a consistency between the majority of observed scores and the expectations of the Rasch model.

Wright map

The Wright maps showed that test items of all versions reflect rather a wide range of difficulty with an even spread. Items are distributed from −2.15 logits (item 16; *SEM* = 0.07) to +2.63 logits (item 65, SEM = 0.18) in IUUESEE 2016, −2.02 logits (item 37; *SEM* = 0.07) to +1.81 logits (item 56, SEM = 0.09) in IUUESEE 2017, −2.21 logits (item 12; *SEM* = 0.06) to +2.15 logits (item 23, SEM = 0.08) in IUUESEE 2018, and −2.55 logits (item 39; *SEM* = 0.07) to +2.98 logits (item 38, SEM = 0.1) in IUUESEE 2019. Furthermore, person ability measures ranged from −4.38 (SEM = 1.88) to +4.69 (SEM = 1.02) in IUUESEE 2016, −3.99 (SEM = 1.86) to +3.79 (SEM = 0.72) in IUUESEE 2017, −5.11 (SEM = 1.89) to +4.2 (SEM = 1.85) in IUUESEE 2018, and −4.9 (SEM = 1.85) to +4.1 (SEM = 0.74) in IUUESEE 2019.

This distribution indicates that test items clustered around the mean, where the majority of test-takers clustered together. Furthermore, no gaps are identified in the item hierarchy. The Maps show similarities between a lot of items in terms of difficulty which denotes the presence of an adequate number of items in the test measuring test-takers' ability, generally near the mean where the majority of test-takers are located. The map also plotted some test-takers above the item with the highest difficulty measure meaning that the tests include some high-ability test-takers whose abilities are beyond the test difficulty.

Rasch reliability analysis

The results of person reliability analyses indicate that 37%, 47%, 51%, and 44% of the variability in person measures of the exams 2016, 2017, 2018, and 2019 respectively are attributable to error. Item reliability estimates show that only 1% of the variability in item measures of test versions 2016, 2017, and 2018 is due to error and there is no sign of error in the variability of item measures of test version 2019.

The person separation of all test versions is around one which refers to the measurement of approximately one statistical strata of performance in persons (Wright, 1996). The analyses of items separation revealed that these measurements consistently measure approximately twelve levels of difficulty in items of exam 2016, ten levels in items of exam 2017, thirteen levels in items of exam 2018, and fifteen levels of difficulty in exam 2019 (Wright, 1996).

Unidimensionality and local independence

We analyzed unidimensionality and local independence with WINSTEPS software. The principal component analysis of linearized Rasch residuals revealed that the Rasch dimension explains 28.1% (eigenvalue=27.4) of observed variance in the exam 2016, 23.5% (eigenvalue=21.4) in the exam 2017, 29.6% (eigenvalue=29.4) in the exam 2018, and 29.7% (eigenvalue=29.5) in the exam 2019 which all variances are remarkably close to the Rasch model prediction of 27.7 %, 23.5%, 29.6%, and 29.7% in the exams 2016, 2017, 2018, and 2019 respectively indicating that the estimation of the Rasch difficulty measures was successful (Linacre, 2010). The first contrast in the residuals explains only 2.3% (eigenvalue=2.2) of the variance in the data in the exam 2016, 2.7% (eigenvalue=2.4) in the exam 2017, 2.2% (eigenvalue=2.2) in the exam 2018, and 2.8% (eigenvalue=2.7) in the exam 2019.

The first extracted dimension from the residuals is about 13.5 times smaller than the Rasch dimension in IUUESEE 2016, 11 times smaller than the Rasch dimension in IUUESEE 2017, 14.5 times smaller than the Rasch dimension in IUUESEE 2018, and 15 times smaller than the Rasch dimension in IUUESEE 2019. Furthermore, the disattenuated correlations of the clusters in all four exams are 1. These statistical outputs support the assumption of unidimensionality in IUUESEE. Investigation of Pearson correlations significantly supported the assumption of local independence. Correlations above 0.70 indicate local dependence (Linacre, 2010), and all observed correlations in the four exams fell between − 0.13 and 0.29 which supported the local independence of all items.

Identification of differential item functioning

IUUESEE 2016

Tables 2, 3, and 4 present native language UNIDIF analysis of test items (items including DIF) of Exam 2016, which include the local difficulty of test items for each native language subgroup, SEM figures for each measurement, the local difficulty contrast between native language subgroups, and a Welch t value and a p value for this contrast. The difference between the local difficulty magnitudes of the items is called the DIF contrast. The Welch t value shows the statistical variance between the local difficulties of items as a Student's double-sided t statistic (Linacre, 2010). For example, Table 3 shows that the difficulty of item 3 is −0.89 with a SEM of 0.09 for the Azeri subgroup and −0.44 with a SEM of 0.17 for the Luri subgroup; the contrast in difficulty, −0.45, is the measure of DIF effect size (Linacre, 2010); the Welch t value of this contrast is −2.39; and the p value of the contrast is 0.0174, which is significant at the established threshold p value of 0.05 indicating that item 3 includes differential functioning based on the

Table 2 Results of Rasch reliability analyses

Test versions	Items reliability	Item separation	Person reliability	Person separation
IUUESEE 2016	0.99	12.38	0.63	1.29
IUUESEE 2017	0.99	10.24	0.53	1.07
IUUESEE 2018	0.99	13.27	0.49	0.98
IUUESEE 2019	1	15.79	0.56	1.26

Table 3 Results of uniform DIF analysis of items (IUUESEE 2016)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Azeri	−0.89	0.09	Luri	−0.44	0.17	−0.45	−2.39	279	0.0174
11	Azeri	−0.22	0.10	Kurdish	−0.70	0.15	0.47	2.61	426	0.0093
23	Azeri	0.48	0.13	Kurdish	−0.01	0.20	0.49	2.04	230	0.0422
23	Azeri	0.48	0.13	Luri	−0.09	0.25	0.57	2.00	120	0.0480
23	Azeri	0.48	0.13	Persian	1.24	0.15	−0.76	−3.85	555	0.0001
28	Azeri	2.02	0.16	Luri	1.12	0.32	0.90	2.51	100	0.0137
36	Azeri	−1.23	0.10	Luri	−0.72	0.20	−0.52	−2.27	184	0.0244
50	Azeri	0.28	0.13	Persian	−0.17	0.13	0.45	2.44	610	0.0150
58	Azeri	0.14	0.17	Persian	−0.35	0.16	0.49	2.18	436	0.0300
61	Azeri	1.13	0.13	Luri	0.00	0.26	1.13	3.87	115	0.0002
62	Azeri	1.60	0.13	Persian	2.03	0.14	−0.43	−2.32	911	0.0207
65	Azeri	2.73	0.28	Luri	1.47	0.52	1.26	2.14	59	0.0365
70	Azeri	0.15	0.15	Persian	−0.37	0.14	0.52	2.55	526	0.0110

Table 4 Results of uniform DIF analysis of items (IUUESEE 2016)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Kurdish	−0.60	0.14	Persian	−1.06	0.09	0.46	2.81	482	0.0051
4	Kurdish	0.54	0.15	Luri	0.04	0.20	0.51	2.02	283	0.0440
14	Kurdish	1.21	0.20	Persian	0.55	0.11	0.65	2.92	290	0.0038
23	Kurdish	−0.01	0.20	Persian	1.24	0.15	−1.25	−4.97	266	0.0000
24	Kurdish	1.27	0.20	Persian	0.71	0.11	0.55	2.37	279	0.0186
29	Kurdish	−1.19	0.13	Persian	−1.70	0.09	0.51	3.15	618	0.0017
50	Kurdish	0.50	0.21	Luri	−0.26	0.28	0.76	2.18	127	0.0313
50	Kurdish	0.50	0.21	Persian	−0.17	0.13	0.67	2.72	217	0.0070
54	Kurdish	−0.09	0.18	Luri	0.51	0.23	−0.60	−2.03	206	0.0440
61	Kurdish	0.87	0.21	Luri	0.00	0.26	0.87	2.60	163	0.0102
66	Kurdish	1.69	0.28	Persian	0.97	0.14	0.72	2.29	152	0.0236

criteria suggested by Linacre (2010). Therefore, it functions differently among the Azeri and the Luri groups based on factors other than the test's construct of interest.

Table 3 presents that uniform differential item functioning analysis identified eleven test items with significant DIF at $p < 0.05$ between the Azeri and the three native language subgroups: items 3, 36, and 62 favoring the Azeri test-takers; item 11 favoring the Kurdish test-takers; items 28, 61, and 65 favoring the Luri test-takers; and items 50, 58, and 70 favoring the Persian test-takers. Items 23 favored the Azeri test-takers compared with the Persian test-takers and the Luri and the Kurdish test-takers in comparison with the Azeri subgroup.

Of these eleven items, four (23, 28, 61, and 65) had UDIF magnitudes larger than 0.6 logits. The item characteristic curves (ICC) of the item with the biggest magnitude (item 65) are presented in Fig. 1. Item 65 favors the Luri test-takers compared with the Azeri ones and leads to differential item functioning.

The solid line in Fig. 1 is the Rasch model curve. Comparing the Azeri with the Luri test-takers, on item 65, the subgroups' ICC curves intersect at four points: − 4.3, −3.6, −3.8, and −1.5 logits (horizontal axis). These are the turning points at which these two

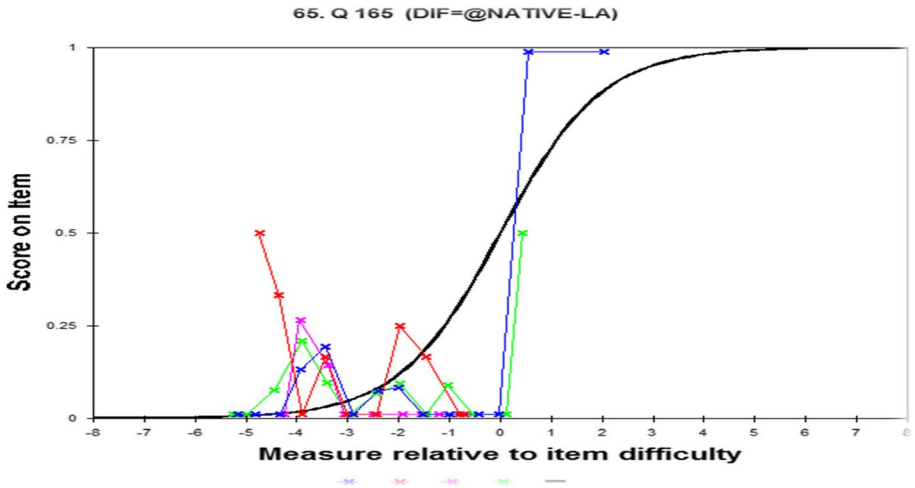

Fig. 1 Item characteristic curves of Item 65 by native language (color figure available online). *Note*: (1) Blue = Azeri. (2) Red = Kurdish. (3) Violet = Lori. (4) Green = Persian

native language subgroups' probabilities of correctly responding to the item intersect. This item favors the Azeri test-takers at ability levels up to − 4.3 logits whereas from − 4.3 to − 3.6 logits, the Luri test-takers were more likely to answer this item correctly; from − 2.8 to − 1.5, the Azeri subgroup was more likely to answer this item correctly. The probability of answering this item at intersecting points was equal for both groups. There was no difference above − 1.5 logits, where many test-takers landed, and from about − 1 to 2 logits where no the Luri takers landed concerning this specific item. The Azeri test-takers were more likely to answer this item correctly.

Table 4 lists ten items with significant DIF at $p < 0.05$ between the Kurdish and the other two native language subgroups namely the Persian and the Luri test-takers. As Table 4 shows, items 23 and 54 are advantageous to the Kurdish test-takers; items 4, 50, and 61 favor the Luri test-takers; and items 3, 14, 24, 29, 50, and 66 favor the Persian test-takers. Six items (14, 23, 50, 54, 61, and 66) of these ten items had UDIF magnitudes larger than 0.6 logits.

According to Table 5, comparing the Luri test-takers with the Persian ones, UDIF analysis identified fifteen test items with significant DIF at $p < 0.05$: items 17, 23, 28, 38, 61, 63, 64, and 65 favoring the Luri test-takers, and items 3, 6, 7, 32, 36, 48, and 57 favoring the Persian test-takers. Of these eleven items, ten (3, 17, 23, 28, 36, 48, 61, 63, 64 and 65) had UDIF magnitudes larger than 0.6 logits.

Figure 2 displays the results of the native language UDIF analysis of all test items for exam 2016. The lines (color figure available online) represent the local item difficulty of the four native language subgroups. The solid line in this figure is the Rasch model curve.

IUUESEE 2017

We explored the native language UDIF analysis for exam 2017. Nineteen DIF items were identified with significant DIF at $p < 0.05$ between the Azeri and the three native language subgroups (Table 6): items 5, 14, 33, 35, 39, 43, and 66 were easier for the Azeri participants; items 3 and 70 functioned in favor of the Persian test-takers; items

Table 5 Results of uniform DIF analysis of items (IUUESEE 2016)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Luri	−0.44	0.17	Persian	−1.06	0.09	0.62	3.30	279	0.0011
6	Luri	0.29	0.18	Persian	−0.17	0.08	0.46	2.34	251	0.0199
7	Luri	−0.80	0.14	Persian	−1.28	0.08	0.48	2.96	390	0.0033
17	Luri	−1.02	0.24	Persian	−0.41	0.12	−0.61	−2.28	133	0.0241
23	Luri	−0.09	0.25	Persian	1.24	0.15	−1.33	−4.51	138	0.0000
28	Luri	1.12	0.32	Persian	2.27	0.16	−1.16	−3.23	100	0.0017
32	Luri	−0.01	0.18	Persian	−0.50	0.09	0.49	2.39	216	0.0177
36	Luri	−0.72	0.20	Persian	−1.35	0.10	0.64	2.78	186	0.0059
38	Luri	0.83	0.24	Persian	1.36	0.12	−0.53	−1.99	169	0.0484
48	Luri	0.05	0.29	Persian	−0.75	0.15	0.81	2.47	99	0.0151
57	Luri	−0.16	0.25	Persian	−0.73	0.13	0.57	2.01	122	0.0467
61	Luri	0.00	0.26	Persian	1.07	0.13	−1.07	−3.67	116	0.0004
63	Luri	−0.57	0.29	Persian	0.19	0.15	−0.76	−2.34	90	0.0217
64	Luri	1.25	0.37	Persian	2.07	0.18	−0.82	−−2.01	86	0.0477
65	Luri	1.47	0.52	Persian	2.93	0.30	−1.46	−2.45	62	0.0172

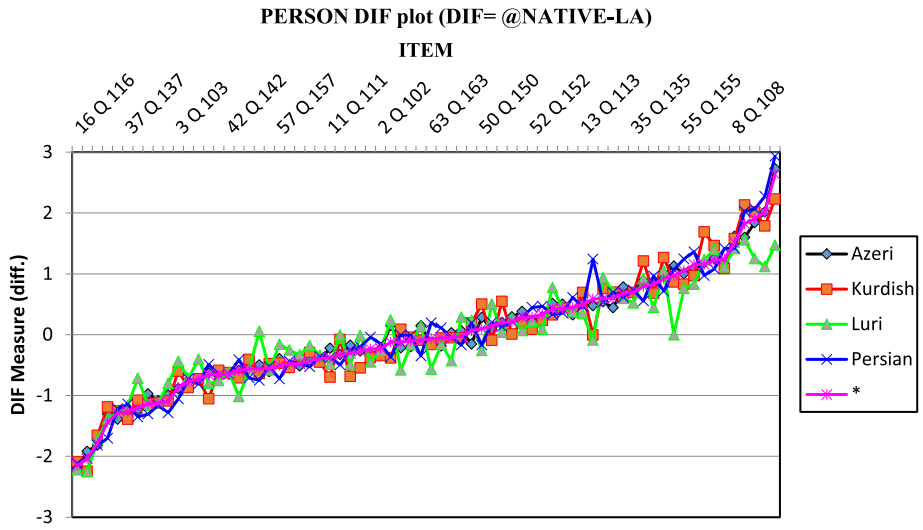

Fig. 2 Uniform differential item functioning in the Iranian Undergraduate University Entrance Special English IUUESEE 2016

15, 24, 53, and 65 were easier for the Kurdish subgroup; and items 8, 12, 16, 24, 25, 30, 56, and 65 favored the Luri test-takers. Standard errors of DIF were not high for all items, although the contrasts were substantive. Of these nineteen items, eleven (12, 15, 16, 24, 25, 30, 35, 56, 65, and 66) had UDIF magnitudes larger than 0.6 logits which indicates that these items were more biased than the items with magnitudes lower than 0.6.

Our native language UNIDIF analysis of the exam 2017 revealed 5 Items with significant DIF at $p < 0.05$ between the Luri and the two native language subgroups: the Persian and the Kurdish (Table 7). All of these items functioned in favor of the Luri test-takers except for item 35 which was easier for the Persian ones. One of these

Table 6 Results of uniform DIF analysis of items (IUUESEE 2017)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Azeri	0.13	0.08	Persian	−0.30	0.08	0.43	3.93	INF	0.0001
5	Azeri	−0.09	0.10	Kurdish	0.48	0.17	−0.57	−2.87	332	0.0044
8	Azeri	0.74	0.10	Luri	0.25	0.21	0.50	2.18	177	0.0308
12	Azeri	0.85	0.13	Luri	0.19	0.22	0.66	2.56	173	0.0112
14	Azeri	−1.33	0.10	Luri	−0.90	0.18	−0.43	−2.09	242	0.0373
15	Azeri	−0.39	0.14	Kurdish	−1.03	0.22	0.64	2.46	204	0.0146
16	Azeri	0.62	0.14	Luri	−0.11	0.27	0.73	2.42	105	0.0172
24	Azeri	1.72	0.16	Kurdish	1.15	0.21	0.57	2.18	351	0.0296
24	Azeri	1.72	0.16	Luri	0.72	0.26	1.00	3.30	164	0.0012
25	Azeri	1.53	0.18	Luri	0.82	0.30	0.71	2.05	142	0.0423
30	Azeri	0.32	0.12	Luri	−0.40	0.23	0.72	2.82	142	0.0055
33	Azeri	−1.45	0.10	Luri	−0.92	0.19	−0.53	−2.52	224	0.0125
35	Azeri	−0.55	0.09	Luri	0.09	0.19	−0.64	−2.99	190	0.0032
39	Azeri	0.64	0.11	Kurdish	1.20	0.18	−0.56	−2.62	369	0.0091
43	Azeri	0.06	0.09	Persian	0.53	0.09	−0.47	−3.57	INF	0.0004
53	Azeri	−0.01	0.11	Kurdish	−0.47	0.17	0.46	2.27	328	0.0239
56	Azeri	2.07	0.16	Luri	1.37	0.29	0.71	2.15	158	0.0331
65	Azeri	1.02	0.15	Kurdish	0.42	0.21	0.60	2.35	234	0.0198
65	Azeri	1.02	0.15	Luri	0.24	0.30	0.78	2.31	83	0.0236
66	Azeri	−0.25	0.16	Luri	0.41	0.34	−0.66	−1.73	70	0.0876
70	Azeri	0.13	0.16	Persian	−0.33	0.15	0.46	2.12	426	0.0342

Table 7 Results of uniform DIF analysis of items (IUUESEE 2017)

Item	*Class A*	*DIF*	*DIF SE*	*Class B*	*DIF*	*DIF SE*	*DIF Contrast*	*Welch t*	*df*	*p*
8	Luri	0.25	0.21	Persian	0.92	0.10	−0.67	−2.94	182	0.0037
12	Luri	0.19	0.22	Kurdish	0.89	0.19	−0.71	−2.40	230	0.0172
12	Luri	0.19	0.22	Persian	1.16	0.13	−0.97	−3.71	182	0.0003
24	Luri	0.72	0.26	Persian	1.84	0.15	−1.12	−3.71	162	0.0003
35	Luri	0.09	0.19	Persian	−0.55	0.09	0.64	3.01	184	0.0030
60	Luri	−0.55	0.28	Persian	0.16	0.14	−0.71	−2.26	94	0.0261

items, item 24 had UDIF magnitudes larger than 0.6 logits which indicates that this item was more biased toward the Luri test-takers than toward the other groups.

Comparing the Kurdish test-takers with the other three native language groups in IUUESEE 2017, our analysis found 6 items with significant DIF at $p < 0.05$ (Table 8). Items 18, 24, 53, and 66 functioned in favor of the Kurdish subgroup, and items 22 and 39 favored the Persian test-takers in comparison with the Kurdish ones. Four items had UNIDIF magnitude larger than 0.6 and item 66 had the largest one leading to more UNIDIF than other items.

The comprehensive results of the native language UDIF analysis of the *IUUESEE 2017* is displayed in Fig. 3.

Table 8 Results of uniform DIF analysis of items (IUUESEE 2017)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
18	Kurdish	−0.55	0.16	Persian	−0.04	0.11	−0.51	−2.67	365	0.0079
22	Kurdish	−1.41	0.14	Persian	−2.00	0.11	0.59	3.25	575	0.0012
24	Kurdish	1.15	0.21	Persian	1.84	0.15	−0.69	−2.65	351	0.0083
39	Kurdish	1.20	0.18	Persian	0.56	0.11	0.64	3.03	359	0.0026
53	Kurdish	−0.47	0.17	Azeri	−0.01	0.11	−0.46	−2.27	328	0.0239
53	Kurdish	−0.47	0.17	Luri	0.21	0.26	−0.68	−2.23	138	0.0273
66	Kurdish	−0.52	0.24	Luri	0.41	0.34	−0.93	−2.22	93	0.0292
66	Kurdish	−0.52	0.24	Persian	0.13	0.16	−0.65	−2.24	153	0.0266

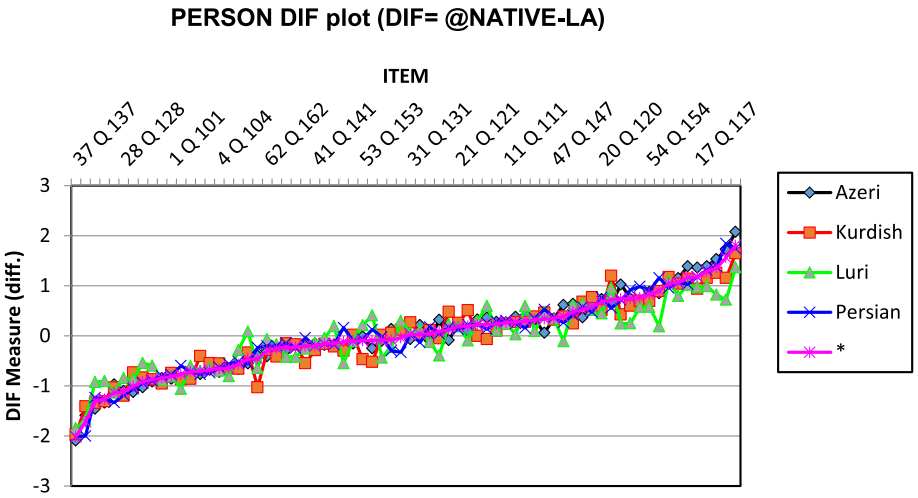

Fig. 3 Uniform differential item functioning in the Iranian Undergraduate University Entrance Special English *IUUESEE 2017*

IUUESEE 2018

The UDIF analysis identified fifteen test items with significant DIF at $p < 0.05$ in IUUESEE 2018 comparing the Azeri test-takers with the Kurdish, the Luri, and the Persian test-takers: items 3, 35, 36, 60, and 61 favoring the Azeri test-takers; items 44, 67, and 70 favoring the Persian test-takers; items 5, 41, 50, 56, 58, and 67 favoring the Luri test-takers; and items 56, 64, and 70 favoring the Kurdish test-takers. Of these eight items, seven items including items 3, 50, 56, 58, 64, 67, and 70 had the UDIF magnitudes larger than 0.60 logits. Item 50 had the largest magnitude and was biased toward the Luri test-takers compared with the Azeri participants (Table 9).

Comparing the Kurdish with the Luri and the Persian test-takers, our analysis revealed 9 items with significant DIF at $p < .05$. Items 3, 30, 36, 64, and 65 were easier for the Kurdish test-takers; items 45, 50, and 58 were easier for the Luri test-takers; and only item 45 functioned in favor of the Persian test-takers. Items 3, 45, 50, 58, 64, and 65 had UDIF magnitude larger than 0.60 (Table 10).

Six items with significant DIF were found in the comparison of the Luri test-takers with the Persian subgroup. Items 41, 50, 58, 65 functioned in favor of the Luri subgroup and items 3 and 4 functioned in favor of the Persian subgroup. Items that were

Table 9 Results of uniform DIF analysis of items (IUUESEE 2018)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Azeri	−0.78	0.07	Luri	0.05	0.15	−0.83	−5.03	359	0.0000
5	Azeri	1.45	0.10	Luri	1.00	0.20	0.46	1.99	257	0.0472
35	Azeri	0.53	0.10	Kurdish	0.99	0.16	−0.46	−2.46	474	0.0144
36	Azeri	−0.32	0.09	Luri	0.16	0.18	−0.48	−2.42	239	0.0165
41	Azeri	1.70	0.13	Luri	1.12	0.26	0.58	2.03	162	0.0444
44	Azeri	0.60	0.09	Persian	0.10	0.09	0.50	3.82	INF	0.0001
50	Azeri	1.83	0.15	Luri	0.95	0.26	0.88	2.95	165	0.0036
56	Azeri	1.31	0.13	Kurdish	0.61	0.18	0.70	3.15	361	0.0018
56	Azeri	1.31	0.13	Luri	0.47	0.24	0.85	3.09	159	0.0024
58	Azeri	−0.35	0.11	Luri	−1.08	0.23	0.73	2.83	140	0.0054
60	Azeri	−0.35	0.11	Kurdish	0.17	0.16	−0.52	−2.69	395	0.0074
61	Azeri	−0.57	0.15	Kurdish	−0.03	0.22	− 0.54	−1.99	203	0.0482
64	Azeri	0.06	0.16	Kurdish	−0.61	0.22	0.67	2.43	211	0.0159
67	Azeri	0.83	0.14	Luri	0.00	0.27	0.83	2.76	112	0.0067
67	Azeri	0.83	0.14	Persian	0.31	0.14	0.53	2.68	555	0.0075
70	Azeri	−0.25	0.14	Kurdish	−0.85	0.22	0.60	2.28	209	0.0239
70	Azeri	−0.25	0.14	Persian	−0.68	0.16	0.43	2.02	447	0.0441

Table 10 Results of uniform DIF analysis of items (IUUESEE 2018)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Kurdish	−0.56	0.11	Luri	0.05	0.15	−0.61	−3.32	480	0.0010
30	Kurdish	0.43	0.21	Luri	0.93	0.35	−0.50	−1.23	89	0.2226
36	Kurdish	−0.31	0.13	Luri	0.16	0.18	−0.47	−2.15	313	0.0325
45	Kurdish	0.58	0.16	Luri	−0.07	0.21	0.65	2.47	224	0.0143
45	Kurdish	0.58	0.16	Persian	0.08	0.10	0.50	2.66	398	0.0080
50	Kurdish	1.83	0.22	Luri	0.95	0.26	0.89	2.63	218	0.0092
58	Kurdish	−0.28	0.16	Luri	−1.08	0.23	0.80	2.84	177	0.0050
64	Kurdish	−0.61	0.22	Persian	0.03	0.16	−0.64	−2.31	214	0.0219
65	Kurdish	0.98	0.30	Persian	1.63	0.24	−0.65	−1.69	153	0.0927

Table 11 Results of uniform DIF analysis of items (IUUESEE 2018)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Luri	0.05	0.15	Persian	−0.51	0.07	0.56	3.46	346	0.0006
4	Luri	0.64	0.16	Persian	0.21	0.07	0.43	2.49	345	0.0131
41	Luri	1.12	0.26	Persian	1.72	0.13	−0.61	−2.11	162	0.0366
50	Luri	0.95	0.26	Persian	1.87	0.14	−0.92	−3.13	156	0.0021
58	Luri	−1.08	0.23	Persian	−0.20	0.11	−0.88	−3.39	139	0.0009
65	Luri	0.61	0.41	Persian	1.63	0.24	−1.03	−2.15	60	0.0360

easier for the Luri test-takers had the UDIF magnitude larger than 0.6 which shows that these items were more biased than items 3 and 4 meaning that this test was more biased toward the Luri test-takers than the Persian ones (Table 11).

Figure 4 illustrates the UDIF analysis of all items of exam 2018.

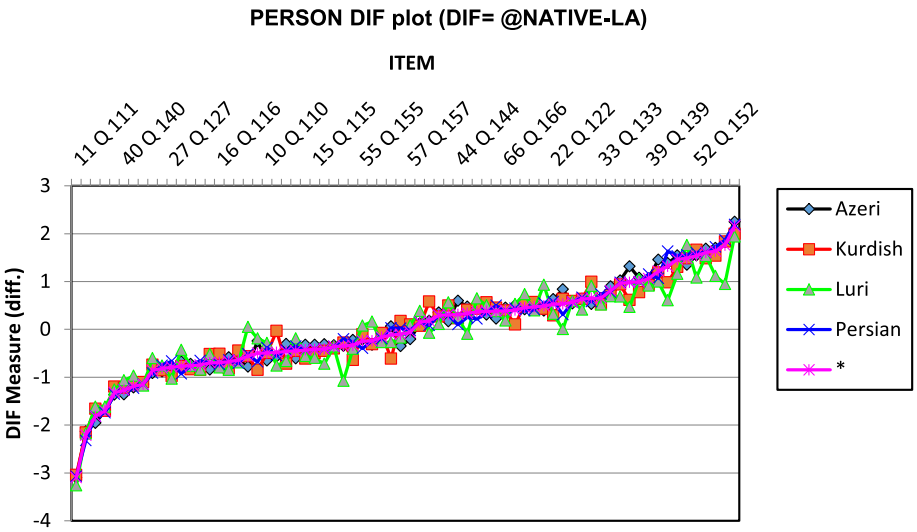

Fig. 4 Uniform differential item functioning in the Iranian Undergraduate University Entrance Special English *IUUESEE 2018*

IUUESEE 2019

The last exam we examined was IUUESEE 2019. Concerning the comparison of the Azeri with the Kurdish, the Persian, and the Luri test-takers, we found 27 items with significant UDIF (Table 12). In this regard, items 11, 14, 15, 20, 22, 37, 38, 39, 40, 42, 51, 65, and 66 functioned in favor of the Azeri test-takers; items 3, 5, 12, 16, 35, 36, 54, 65, and 70 favored the Luri test-takers; items 21, 27, 35, 55, 62, and 68 were easier for the Kurdish test-takers; and items 35 and 50 were easier for the Persian test-takers. Items 35, 40, 65, 66, and 68 had UDIF magnitude larger than 0.6 indicating strong construct-irrelevant variance.

Table 13 shows the eighteen test items with significant DIF comparing the Kurdish test-takers with the Luri and the Persian ones in IUUESEE 2019. Items 11, 22, 27, 37, 39, 40, 43, 59, and 68 were easier for the Kurdish test-takers; items 12, 14, 16, 35, 65, and 70 functioned in favor of the Luri test-takers; and items 14, 20, 42, and 57 were easier for the Persian test-takers compared with the Kurdish subgroup. Items 12, 22, 37, 40, 59, 65, 68, and 70 had the UDIF magnitude larger than 0.6 from which item 65 had the largest one which was biased toward the Luri test-takers like most of the items having the largest magnitudes.

Comparing the Luri with the Persian test-takers, we found 12 items with a significant DIF (Table 14). Most of these items including items 5, 16, 19, 22, 25, 45, 55, and 70 functioned in favor of the Luri test-takers and only items 15, 40, 44, and 66 were easier for the Persian test-takers. Items 19, 22, 40, 66, and 70 had larger magnitudes than 0.6. Item 70 had the largest UDIF magnitude meaning that this item was biased toward the Luri test-takers as was the case with most of the items with the largest magnitudes.

Table 15 demonstrates the overall number of UDIF instances in each test version. From 280 items in four test versions and 24 comparisons, a total of 165 instances of significant UDIF at $p < 0.05$ were detected:

Table 12 Results of uniform DIF analysis of items (IUUESEE 2019)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
3	Azeri	1.18	0.13	Luri	0.61	0.24	0.57	2.13	183	0.0345
5	Azeri	−1.86	0.08	Luri	−2.29	0.17	0.43	2.31	414	0.0212
11	Azeri	0.32	0.08	Luri	0.78	0.17	−0.46	−2.42	307	0.0160
12	Azeri	−1.93	0.10	Luri	−2.40	0.21	0.47	2.06	274	0.0406
14	Azeri	−0.32	0.09	Kurdish	0.18	0.15	−0.50	−2.87	405	0.0043
15	Azeri	−0.43	0.08	Luri	0.11	0.17	−0.54	−2.86	240	0.0046
16	Azeri	0.27	0.08	Luri	−0.28	0.15	0.55	3.15	307	0.0018
20	Azeri	−1.72	0.08	Kurdish	−1.18	0.12	−0.54	−3.66	692	0.0003
21	Azeri	2.35	0.15	Kurdish	1.82	0.22	0.53	1.98	415	0.0482
22	Azeri	0.50	0.14	Persian	1.03	0.13	−0.52	−2.72	582	0.0067
27	Azeri	0.51	0.09	Kurdish	−0.01	0.15	0.52	2.96	380	0.0033
35	Azeri	1.36	0.11	Kurdish	0.87	0.17	0.49	2.46	427	0.0142
35	Azeri	1.36	0.11	Luri	0.32	0.20	1.04	4.64	229	0.0000
35	Azeri	1.36	0.11	Persian	0.66	0.09	0.70	4.95	INF	0.0000
36	Azeri	−0.36	0.08	Luri	−0.87	0.16	0.51	2.79	259	0.0057
37	Azeri	−0.20	0.08	Luri	0.30	0.18	−0.50	−2.50	231	0.0130
38	Azeri	2.84	0.16	Persian	3.31	0.18	−0.47	−1.99	INF	0.0466
39	Azeri	−2.63	0.11	Luri	−2.15	0.18	−0.48	−2.25	405	0.0251
40	Azeri	−0.94	0.08	Luri	−0.22	0.15	−0.72	−4.21	300	0.0000
42	Azeri	−0.88	0.09	Kurdish	−0.38	0.15	−0.50	−2.86	384	0.0045
50	Azeri	−0.97	0.12	Persian	−1.54	0.14	0.57	3.08	724	0.0022
51	Azeri	−1.18	0.10	Luri	−0.74	0.20	−0.44	−1.97	175	0.0500
54	Azeri	0.93	0.13	Luri	0.39	0.28	0.54	1.74	99	0.0843
55	Azeri	1.52	0.13	Kurdish	0.94	0.19	0.58	2.47	333	0.0140
62	Azeri	0.07	0.16	Kurdish	−0.37	0.22	0.44	1.60	194	0.1115
65	Azeri	0.92	0.19	Kurdish	1.38	0.33	−0.46	−1.22	122	0.2260
65	Azeri	0.92	0.19	Luri	−0.05	0.43	0.97	2.04	38	0.0484
66	Azeri	0.23	0.12	Luri	0.92	0.32	−0.68	−1.99	98	0.0494
68	Azeri	2.85	0.21	Kurdish	2.05	0.26	0.81	2.44	371	0.0151
70	Azeri	1.27	0.20	Luri	0.02	0.36	1.25	3.01	69	0.0036

Non-uniform differential item functioning

To conduct a NUDIF analysis of the four versions of *IUUESEE*, we segmented native language groups into high- and low-ability subgroups by partitioning the range of person ability measures at the point in the middle of the range and then performed a NUDIF analysis of all test items. In this regard, WINSTEPS invoked 7840 NUDIF comparisons for 280 test items considering eight native language subgroups. When high-ability and low-ability subgroups were compared, a total of 1730 instances of significant NUDIF at $p < 0.05$ were revealed (Table 16).

Our analysis found that IUUESEE 2019 had the largest number of NUDIF cases and IUUESEE 2017 the fewest. The largest number of NUDIF cases was found to relate to the low-ability Persian test-takers and the fewest number of NUDIF instances dealt with the low-ability Luri test-takers. In this case, the largest number of NUDIF cases including 515 instances dealt with the Persian test-takers, and the fewest numbers of NUDIF cases including 299 instances related to the Luri test-takers.

Table 13 Results of uniform DIF analysis of items (IUUESEE 2019)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
11	Kurdish	0.29	0.13	Luri	0.78	0.17	−0.49	−2.25	435	0.0252
12	Kurdish	−1.75	0.15	Luri	−2.40	0.21	0.64	2.54	358	0.0115
14	Kurdish	0.18	0.15	Luri	−0.34	0.17	0.52	2.29	352	0.0227
14	Kurdish	0.18	0.15	Persian	−0.31	0.09	0.49	2.85	395	0.0046
16	Kurdish	0.16	0.13	Luri	−0.28	0.15	0.44	2.16	425	0.0310
20	Kurdish	−1.18	0.12	Persian	−1.70	0.08	0.52	3.54	677	0.0004
22	Kurdish	0.41	0.24	Persian	1.03	0.13	−0.62	−2.24	156	0.0262
27	Kurdish	−0.01	0.15	Persian	0.44	0.09	−0.44	−2.53	375	0.0118
35	Kurdish	0.87	0.17	Luri	0.32	0.20	0.55	2.15	306	0.0322
37	Kurdish	−0.32	0.13	Luri	0.30	0.18	−0.62	−2.76	325	0.0061
39	Kurdish	−2.72	0.18	Luri	−2.15	0.18	−0.57	−2.26	528	0.0241
40	Kurdish	−1.02	0.12	Luri	−0.22	0.15	−0.80	−4.09	426	0.0001
42	Kurdish	−0.38	0.15	Persian	−0.94	0.10	0.56	3.11	408	0.0020
43	Kurdish	0.36	0.13	Luri	0.88	0.19	−0.51	−2.19	351	0.0290
57	Kurdish	1.00	0.18	Persian	0.42	0.11	0.58	2.74	303	0.0065
59	Kurdish	−0.69	0.24	Persian	−0.01	0.15	−0.68	−2.39	156	0.0179
65	Kurdish	1.38	0.33	Luri	−0.05	0.43	1.43	2.63	59	0.0110
68	Kurdish	2.05	0.26	Persian	2.70	0.19	−0.65	−2.05	330	0.0407
70	Kurdish	0.95	0.30	Luri	0.02	0.36	0.93	1.99	93	0.0499

Table 14 Results of uniform DIF analysis of items (IUUESEE 2019)

Item	Class A	DIF	DIF SE	Class B	DIF	DIF SE	DIF Contrast	Welch t	df	*p*
5	Luri	−2.29	0.17	Persian	−1.79	0.08	−0.50	−2.72	404	0.0068
15	Luri	0.11	0.17	Persian	−0.33	0.08	0.45	2.37	235	0.0186
16	Luri	−0.28	0.15	Persian	0.16	0.08	−0.44	−2.54	304	0.0116
19	Luri	0.80	0.18	Persian	1.53	0.10	−0.73	−3.60	328	0.0004
22	Luri	0.34	0.32	Persian	1.03	0.13	−0.69	−1.97	75	0.0523
25	Luri	0.18	0.23	Persian	0.72	0.12	−0.54	−2.09	147	0.0386
40	Luri	−0.22	0.15	Persian	−0.86	0.07	0.64	3.77	295	0.0002
44	Luri	0.01	0.19	Persian	−0.09	0.09	0.10	0.49	209	0.6268
45	Luri	1.09	0.29	Persian	1.57	0.13	−0.48	−1.52	133	0.1301
55	Luri	1.06	0.28	Persian	1.31	0.12	−0.25	−0.83	128	0.4077
66	Luri	0.92	0.32	Persian	−0.17	0.13	1.08	3.14	101	0.0022
70	Luri	0.02	0.36	Persian	1.19	0.19	−1.17	−2.86	65	0.0056

Table 15 Number of UDIF cases favoring each group in each test version

Native language groups	Test versions				Total
	2016	2017	2018	2019	
Azeri	4	7	5	13	29
Persian	16	5	6	10	37
Kurdish	4	8	8	15	35
Luri	15	13	13	23	64
Total	39	33	32	61	165

Table 16 Number of NUDIF cases favoring each subgroup in each test version

Native language subclasses	Test versions				Total
	2016	2017	2018	2019	
Azeri 1	55	48	48	67	218
Azeri 2	84	66	64	78	292
Persian 1	56	45	56	58	215
Persian 2	90	54	71	85	300
Kurdish 1	50	55	51	65	221
Kurdish 2	47	39	52	47	185
Luri 1	53	60	64	59	236
Luri 2	16	16	24	7	63
Total	451	383	430	466	1730

Overall, from 7840 possible comparisons between eight native language subgroups, 22% of NUDIF instances were revealed and from 1960 possible comparisons in each test version, test versions 2016, 2017, 2018, and 2019 respectively included 23%, 19.5%, 21.9%, and 23.7% of NUDIF instances.

Discussion

This study set out to investigate DIF caused by native language in Undergraduate University Entrance Special English Exam (IUUESEE), using the logistic Rasch model. Overall, the results of this study showed that item format and content of the IUUESEE interact with the native language of test-takers and form bias in the evaluation of their performance. Analysis of descriptive statistics, item difficulty measures, fit to the Rasch model, unidimensionality, local independence, and reliability fulfilled the requirements for DIF analysis.

Reliability analysis of IUUESEE found strong support for the item reliability and separation of the test; however, it cast doubt on the person's reliability and separation of test items based on Linacre (2012). The findings showed that IUUESEE resulted in a lower ability range and has not probably distinguished between high performers low performers appropriately (Linacre, 2012). Our DIF and fit results support this finding. On the other hand, high item reliability and separation coefficients of IUUESEE indicate that it measured the wide range of difficulty and also our sample was large enough to accurately locate the items on the latent variable (Linacre, 2012).

Our investigation of Pearson correlations supported the local independence of items and dimensionality and PCAR analyses revealed that test-takers' performances are not influenced by off-dimensional components to a considerable extent. The test items have not constructed different patterns or clusters which supports unidimensionality (Linacre, 2010).

Fit analysis of the four test visions satisfied the preconditions for DIF analysis. Fit indices of the majority of the items were 1 or near 1 which indicated a lack of erratic response patterns in the data. Although due to the lack of conventional Rasch fit criteria, we were not certain whether Bond and Fox's (2007) more lenient criterion operated better than Wright and Linacre's (1994) more rigid one or not, our findings showed a few erratic response patterns across the data based on Wright and Linacre's criterion. Furthermore, it was also found that Wright and Linacre's (1994) fit criterion (0.8–1.2) was more advantageous than other criteria such as Bond and Fox's criterion (2007) in the investigation of test-takers' response patterns.

Overall, the majority of the items across all test versions, such as items 3 and 12 in test version 2017, showed MNSQ fit indices of 1 or near 1 which suggested an absence of erratic response patterns in the data. However, IUUESEE also included several misfitting items. Several items such as items 7 and 29 in version 2016, items 9 and 10 in version 2017, items 2 and 17 in version 2018, and items 15 and 26 in version 2019 showed MNSQ fit indices below 1 and overfit the model, and some items, such as items 15 and 20 in version 2016, items 16 and 18 in version 2017, items 23 and 27 in version 2018, and items 19 and 22 in version 2019 underfit the model to some extent, leading to unexpected variance which is likely due to carelessness or guessing (Wright & Linacre, 1994).

Misfitting items of The IUUESEE do not provide test-takers with equal opportunities to demonstrate their language proficiency. These items overestimate test-takers who could function worse and underestimate those who could function better undermining the fairness of the test. In the case of the easiest misfitting items which their outfit MNSQ values misfit due to sensitivity to outliers, high-ability test-takers missed these easy items. Concerning the most difficult misfitting items with sensitivity to outliers, test-takers with lower levels of language proficiency answered these difficult items correctly. For instance, items 11 and 12 were the easiest test items of test version 2018 (11 difficulty measure= −3.1; 12 difficulty measure= -2.21). Their outfit MNSQ values misfit (11 outfit MNSQ=0.67; 12 outfit MNSQ=0.75). Because outfit is sensitive to outliers, this shows that some high-ability test-takers missed these easy items (Bond & Fox, 2007). The outfit MNSQ values of items 28 (difficulty measure= 2.02) and 65 (difficulty measure= 2.63) which were the most difficult items of test version 2016 were 2.14 and 3.28 respectively indicating that low-ability test-takers answered these difficult items correctly. This finding indicates that determining more strict fit criteria in Rasch-based analysis of dichotomous data contributes to the identification of erratic patterns which is likely attributable to a perplexing impact on an item-level such as DIF (Smith, 1996). That is why we used Wright and Linacre's (1994) range from 0.8 to 1.2 in this study.

Based on UDIF analysis, we explored that IUUESEE 2019 had the largest instances of significant UDIF (61 cases) which is consistent with our findings of NUDIF analysis. It was found that, in most cases (64), items functioned in favor of the Luri test-takers compared to test-takers from other native language groups. Azeri test-takers were favored on the smallest number of items displaying UDIF. NUDIF analysis revealed that a large number of NUDIF instances have happened in favor of the low-ability Persian, the low-ability Azeri, the high-ability Kurdish, and the high-ability Luri test-takers. Finding the real sources of observed DIF is often demanding (Camilli & Shepard, 1994; Gierl, 2005), especially in exploratory DIF investigations which lack a priori hypothesis (Jang & Roussos, 2009); however, reviewing items provided us with some reasons. It showed that low-ability and the Luri test-takers had answered several difficult misfitting items correctly which their counterparts missed. This assumption is supported by the outfit MNSQ patterns of these items: several correct answers on difficult test items by low-ability and the Luri test-takers had outfit MNSQ values greater than 1.2 and lower than 0.8 indicating that their performance on these items was unexpected and can be related most likely to successful lucky guesses. Since all items of the test versions were multiple choice having four options, attempting a lucky guess has a chance of success (25 %).

A closer look at items revealed that the participants had to match the given source text with the sentences which were paraphrased. This entailed a higher level of comprehension. It appears that in these types of items readers needed to establish a text base understanding and keep it in their memory to form a position model which integrates the new coming information with the surface information (Kintsch, 1998). According to Kintsch (1998), this surface information needs to contain a robust mental representation of the elements of the two passages that the test-takers need to match against the mental representation of the correct choice by higher-level cognitive processing. The extent to which a test-takers performs this cognitive processing successfully determines whether the test-takers could answer a test item correctly, and the complexity of this process may encourage guessing from low-level test-takers who cannot successfully carry out the comprehension process (Xuelian & Aryadoust, 2020).

The presence of wrong answers in the responses of the high-ability test-takers supports the assumption that they did not answer easy items correctly probably due to carelessness, overconfidence, and thoughtless errors (Aryadoust et al., 2011). This assumption is also supported by results of the fit analysis which revealed that high-ability test-takers missed the easiest misfitting items which their outfit MNSQ values misfit due to sensitivity to outliers (extreme values) in the data set.

Socioeconomic status has a significant role in test-takers' performance (see, e.g., Şirin, 2005; Suleman et al., 2012; Kormos & Kiddle, 2013). It may be another source of DIF across native language groups. A factor relating to the socio-economic status that might have influenced item performance in IUUESEE is test-wiseness. It refers to the familiarity with the test format, because of test-takers' educational background, which can affect test performance (Xuelian & Aryadoust, 2020). Two hundred ninety-two and 300 instances of NUDIF occurred in favor of the low-ability Azeri and the low-ability Persian test-takers respectively. This finding points to the importance of test-wiseness and its effect on test-takers' performance. Since the Azeri and the low-ability Persian test-takers were mostly from high and middle socio-economic areas, they had afforded to participate in IUUESEE preparation courses and equip themselves with test-taking strategies to succeed in the test. As Hayes and Read (2004) stated, test-takers with previous exposure to an exam are trained in specific test-taking strategies to respond to test items, which might have assisted them to answer items that their counterparts could not. This echoes the remarks of Ryan and Bachman (1992) who stated that language background "is most likely a surrogate for a complex of cultural, societal, and educational differences" (p. 11). Therefore, the native language may be considered as a representative factor that causes DIF in items, rather than the main source of the DIF (Xuelian & Aryadoust, 2020).

The finding that the large number of items displaying UDIF favored the Luri test-takers is in contrast to our expectation based on their socioeconomic status, since the majority of the Luri test-takers are from low socio-economic areas in Iran (Chalabi & Janadele, 2007). This is supported by the results of the fit analyses (see Tables 17, 18, 19, and 20 in the Appendix) and the general picture of native language-based UDIFanalyses (see Figs. 2, 3, 4, and 5). The figures show that the Luri native language group deviates more from the Rasch model curve than the other groups refering to the largest number of the Luri test-takers' erratic response patterns in the data as was found by fit analyses. Moreover, the figures indicate that the Azeri native language group included the

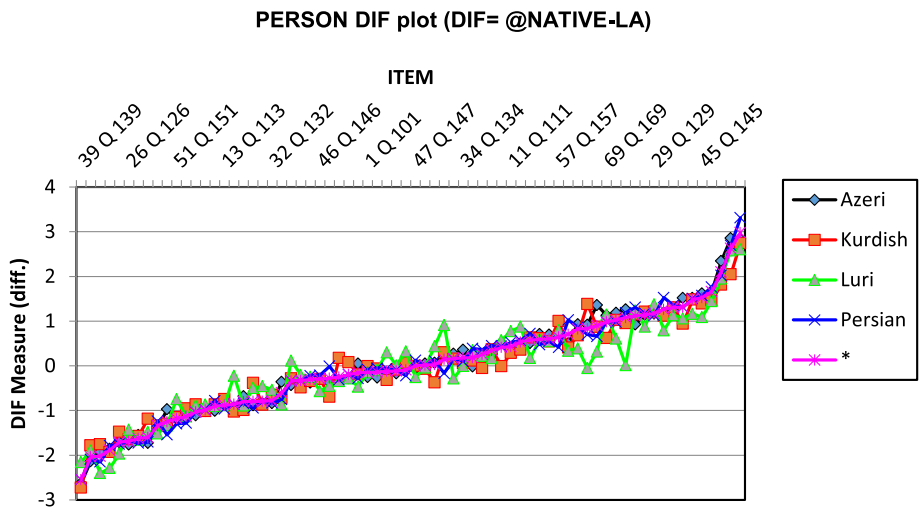

Fig. 5 Uniform differential item functioning in the Iranian Undergraduate University Entrance Special English *IUUESEE 2019*

smallest number of erratic response patterns across the four versions of IUUESEE which was supported by specific UDIF results. In general, we should acknowledge that, in this study, the native language is most likely a surrogate for a combination of educational, cultural, and societal differences (Ryan & Bachman, 1992) and the reasons for the DIF may simultaneously derive from different sources and in some contexts, they may not be so obvious (Schmitt et al., 1993).

Finally, we should not ignore the effect of the test-takers' native language on their test performance from a developmental, second language acquisition (SLA) perspective. Due to the effect of Ll on L2 acquisition, considering DIF as a function of Ll in IUUESEE is not surprising. The majority of SLA researchers have reached a consensus that the effect of Ll is mostly greatest at the initial stages of SLA, or at the lower levels of L2 proficiency, and is likely to decrease as L2 proficiency is increased, leading to greater DIF at lower L2 ability levels and less DIF at higher ability levels (Ryan & Bachman, 1992). Bradlow and Bent (2008) found that the native language effect was less obvious as language proficiency developed and advanced. The findings of the current study resonated with Bradlow and Bent's (2008) findings.

Therefore, in IUUESEE, we need to acknowledge the sensitivity of DIF to the low-ability test-takers, since test-takers' native language influenced their test performance as a result of the insufficient development of their target language. This is in line with the results of another study which found that the dimensionality of L2 exams is a function of test-takers' proficiency levels (Oltman et al., 1988).

Conclusion and future research

This study has provided insight into the interaction between test-takers' native language and their test performance. This interaction became more evident when native language groups were divided into subgroups in NUDIF analysis. The UDIF and the NUDIF analyses respectively revealed 165 and 1730 instances of significant DIF at the established threshold *p*-value of 0.05 recommended by Linacre (2010). It was found that the Luri test-takers were favored more on the test items of IUUESEE than other native language groups. Since the IUUESEE is an MC format, it appears that it encouraged lucky guesses

among the Luri test-takers, the low-ability Azeri, and the low-ability Persian test-takers who have probably practiced test-taking strategies. Closer inspection of the DIF items showed that many of them had long and wordy stems and unappealing distractors. These factors, together with a little time available to answer the test items probably led low-ability and the Luri test-takers to venture lucky guesses. This was also supported by the results of fit analyses which showed erratic response patterns in the data and revealed that low-ability test-takers have successfully answered difficult items with misfitting outfit MNSQ values due to sensitivity to outliers. Test-takers' socioeconomic status appeared to be another factor contributing to DIF results.

These findings suggest some implications for the Iranian National University Entrance English Exam. The findings cast doubt on the validity of the IUUESEE by examining its item functionality across different native language groups and subgroups of test-takers. This information is useful for stakeholders such as test writers and policymakers. They should be cognizant of the issue that some items of IUUESEE display DIF among test-takers with different native languages leading to construct-irrelevant variance. Deciding whether to keep or eliminate DIF items would entail an examination of the whole test and would depend on the application of the cancellation rule (Borsboom, 2006). However, the test designers need to inspect the test bank to know whether there are similar DIF items and offer transparent procedures for item writers to eschew systematic problems. Considering this perspective, the current study offers empirical evidence that can be put into consideration for improving the design of IUUESEE.

There are some limitations in this study that should be stated. Our study is limited in scope as it examined native language as a separate factor leading to DIF. However, as we mentioned, there are other factors including socioeconomic status (e.g., test preparation), guessing, overconfidence, thoughtless errors, stem length, time, and unappealing distractors which may be sources of DIF in IUUESEE. These factors need to be investigated meticulously to help test designers better evaluate their effect on IUUESEE outcomes. Future research also needs to investigate other aspects, such as age, content and item type, academic background, and prior exposure to English, which have all been revealed as sources of DIF in previous studies (Aryadoust, 2012; Chubbuck et al., 2016; Pae, 2004; Takala & Kaftandjieva, 2000). Recent developments in latent DIF analysis that integrated Rasch measurement with latent class analysis can pave the way for future research and address the complications in DIF research with manifest variables (Benıtez et al., 2016; Cohen & Bolt, 2005; Strobl et al., 2015). According to Zumbo's (2007) third generation of DIF, examination of socio-cultural and contextual factors which may affect different native language groups' performances differently would be an interesting domain of investigation. Since task types and test content are undoubtedly the main determinants of test-takers' test performances, we believe that one line of inquiry for continued research would be quantitative analyses across task types and qualitative studies examining the content of the test with a panel of experts which can shed light on the relationships between item content and DIF.

Another limitation relates to the nature of the exploratory DIF approach upon which our study is grounded. Although this approach was able to find UDIF and NUDIF in some items, it failed to find and clarify the causes of DIF. Future research should explore possible ways to perform confirmatory native language-based DIF study of IUUESEE to unravel the DIF sources (e.g., Gierl, 2005).

Appendix

Tables 17,18,19 and 20

Table 17 Results of descriptive statistics analysis and Rasch measurement (IUUESEE 2016)

Descriptive statistics						*Rasch measurement*			
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
1	0.58	0.49	−0.36	−1.87	−0.66	1.03	1.04	0.41	1323
2	0.47	0.49	0.08	−1.99	−0.23	0.92	0.90	0.51	1131
3	0.63	0.48	−0.54	−1.70	−0.86	0.93	0.93	0.48	1200
4	0.43	0.49	0.25	−1.93	0.02	0.84	0.84	0.56	716
5	0.59	0.49	−0.37	−1.86	−0.78	1.09	1.31	0.33	709
6	0.46	0.49	0.15	−1.97	−0.03	0.94	0.91	0.49	879
7	0.66	0.47	−0.71	−1.49	−1.09	0.86	0.78	0.54	1664
8	0.20	0.40	1.47	0.16	1.51	0.98	1.10	0.38	333
9	0.42	0.49	0.30	−1.90	0.33	1.21	1.26	0.25	560
10	0.36	0.48	0.57	−1.67	0.60	0.92	0.92	0.50	524
11	0.54	0.49	−0.16	−1.97	−0.38	0.95	0.99	0.48	710
12	0.25	0.43	1.12	−0.72	1.24	1.05	1.21	0.36	288
13	0.35	0.47	0.61	−1.62	0.50	0.88	0.86	0.53	670
14	0.31	0.46	0.80	−1.35	0.80	0.91	0.94	0.51	379
15	0.43	0.49	0.27	−1.93	0.29	1.24	1.34	0.34	272
16	0.83	0.36	−1.84	1.42	−2.15	1.00	1.08	0.33	1618
17	0.62	0.48	−0.51	−1.74	−0.59	0.97	0.96	0.46	666
18	0.60	0.48	−0.44	−1.80	−0.56	0.99	0.96	0.43	879
19	0.54	0.49	−0.16	−1.97	−0.31	1.05	1.00	0.44	433
20	0.41	0.49	0.36	−1.87	0.43	1.21	1.27	0.32	344
21	0.83	0.37	−1.77	1.13	−2.04	0.91	0.92	0.41	1649
22	0.82	0.37	−1.72	0.96	−1.78	0.78	0.59	0.59	1176
23	0.37	0.48	0.53	−1.71	0.59	1.31	1.57	0.21	298
24	0.28	0.45	0.94	−1.11	0.92	0.87	0.91	0.52	357
25	0.71	0.44	−0.97	−1.05	−1.29	0.89	0.86	0.49	1557
26	0.57	0.49	−0.28	−1.91	−0.47	0.96	0.93	0.46	1105
27	0.31	0.46	0.78	−1.38	0.83	1.34	1.61	0.07	609
28	0.16	0.37	1.78	1.19	2.02	1.32	2.42	0.05	155
29	0.75	0.43	−1.15	−0.66	−1.44	0.86	0.76	0.50	1617
30	0.61	0.48	−0.48	−1.76	−0.43	0.87	0.82	0.54	704
31	0.64	0.47	−0.61	−1.62	−0.73	0.88	0.82	0.51	1079
32	0.54	0.49	−0.18	−1.96	−0.33	0.87	0.85	0.54	888
33	0.43	0.49	0.25	−1.93	0.21	1.01	1.03	0.42	700
34	0.69	0.46	−0.84	−1.28	−1.12	0.99	1.06	0.39	1283
35	0.34	0.47	0.66	−1.56	0.69	1.00	1.05	0.41	523
36	0.73	0.44	−1.07	−0.83	−1.21	0.88	0.80	0.49	1133
37	0.73	0.44	−1.05	−0.88	−1.26	0.93	0.85	0.45	1236
38	0.24	0.43	1.16	−0.63	1.15	1.08	1.23	0.35	329
39	0.49	0.50	0.005	−2.0	−0.06	1.07	1.06	0.35	788
40	0.51	0.50	−0.04	−2.0	−0.10	1.09	1.10	0.36	710
41	0.54	0.49	−0.18	−1.96	−0.39	1.03	1.02	0.39	1131
42	0.61	0.49	−0.48	−1.77	−0.65	1.05	1.04	0.37	1084
43	0.40	0.49	0.38	−1.85	0.44	1.08	1.11	0.34	705
44	0.39	0.48	0.41	−1.85	0.43	1.01	1.04	0.41	610

Table 17 (continued)

Descriptive statistics						*Rasch measurement*			
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
45	0.55	0.49	−0.22	−1.95	−0.27	1.02	1.10	0.40	676
46	0.60	0.49	0.41	−1.82	−0.48	0.99	0.97	0.41	904
47	0.37	0.48	0.52	−1.72	0.68	1.09	1.15	0.34	461
48	0.65	0.47	−0.63	−1.59	−0.56	0.78	0.72	0.62	472
49	0.61	0.48	0.48	0.48	−0.56	0.94	0.94	0.46	910
50	0.51	0.50	0.06	−2.0	0.10	0.84	0.81	0.59	410
51	0.51	0.50	−0.04	−2.00	0.06	0.94	0.92	0.48	620
52	0.45	0.49	0.18	−1.96	0.29	1.08	1.09	0.34	686
53	0.41	0.49	0.35	−1.87	0.60	1.10	1.16	0.36	399
54	0.49	0.50	0.03	−2.00	0.14	1.10	1.09	0.36	534
55	0.30	0.46	0.83	−1.30	1.05	1.08	1.23	0.35	329
56	0.54	0.49	−0.18	−1.96	−0.24	1.18	1.21	0.29	700
57	0.63	0.48	−0.57	−1.67	−0.51	0.88	0.79	0.54	598
58	0.56	0.49	−0.25	−1.94	−0.09	0.90	0.85	0.58	326
59	0.27	0.44	1.03	−0.93	1.22	0.83	0.82	0.55	340
60	0.53	0.49	−0.15	−1.98	−0.12	1.00	1.05	0.53	253
61	0.32	0.46	0.75	−1.43	0.97	1.48	1.78	0.04	279
62	0.17	0.38	1.70	0.92	1.82	1.17	1.43	0.21	213
63	0.53	0.49	−0.15	−1.9	−0.06	1.05	1.07	0.43	345
64	0.18	0.18	0.18	0.18	1.91	1.08	1.68	0.29	126
65	0.09	0.29	2.76	5.65	2.63	1.37	3.28	−0.05	42
66	0.31	0.46	0.77	−1.39	1.16	1.10	1.42	0.34	214
67	0.75	0.42	0.42	0.42	−1.14	0.81	0.70	0.56	692
68	0.54	0.49	−0.17	−1.97	−0.05	0.87	0.83	0.55	620
69	0.69	0.46	−0.84	−1.28	−0.66	0.91	0.88	0.52	470
70	0.57	0.49	−0.28	−1.92	−0.12	0.90	0.88	0.56	387

Table 18 Results of descriptive statistics analysis and Rasch measurement (IUUESEE 2017)

Descriptive statistics						*Rasch measurement*			
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
1	0.59	0.49	−0.37	−1.86	−0.81	0.85	0.79	0.55	962
2	0.39	0.48	0.42	−1.82	0.04	1.06	1.11	0.33	739
3	0.42	0.49	0.28	−1.91	−0.08	1.04	1.04	0.36	912
4	0.55	0.49	−0.23	−1.94	−0.66	0.97	0.95	0.42	1010
5	0.41	0.49	0.32	−1.89	0.12	0.95	0.96	0.44	508
6	0.53	0.49	−0.12	−1.9	−0.57	0.87	0.84	0.51	1142
7	0.20	0.40	1.49	0.22	1.28	1.06	1.26	0.25	278
8	0.26	0.44	1.06	0.87	0.74	1.06	1.13	0.29	441
9	0.41	0.49	0.33	−1.89	0.03	0.99	0.99	0.40	674
10	0.44	0.49	0.22	−1.95	−0.15	0.98	0.98	0.40	883
11	0.36	0.48	0.57	−1.66	0.27	0.87	0.87	0.53	429
12	0.25	0.43	1.13	−0.71	0.89	1.01	1.09	0.35	261
13	0.23	0.42	1.26	−0.39	1.04	1.02	1.08	0.34	258
14	0.67	0.46	−0.74	−1.44	−1.27	0.86	0.78	0.52	1094
15	0.54	0.49	−0.17	−1.97	−0.47	1.00	0.98	0.42	364
16	0.36	0.48	0.56	−1.68	0.38	1.24	1.29	0.18	286
17	0.18	0.38	1.62	0.63	1.18	0.98	1.13	0.34	186
18	0.49	0.50	0.01	−2.00	−0.20	1.19	1.27	0.20	539
19	0.40	0.49	0.38	−1.85	0.19	1.08	1.13	0.35	319
20	0.30	0.46	0.46	0.46	0.68	1.07	1.16	0.30	400
21	0.39	0.48	0.42	−1.82	0.19	0.92	0.90	0.49	433
22	0.77	0.41	−1.31	−0.27	−1.71	0.91	0.79	0.44	1245
23	0.66	0.47	−0.68	−1.53	−0.92	0.86	0.79	0.53	588
24	0.16	0.37	1.77	1.15	1.56	1.22	1.61	0.13	171
25	0.19	0.39	1.50	0.27	1.35	1.19	1.51	0.20	138
26	0.38	0.48	0.46	−1.79	0.24	1.20	1.23	0.18	533
27	0.64	0.47	−0.58	−1.65	−1.01	0.96	0.94	0.40	949
28	0.66	0.47	−0.68	−1.53	−1.13	0.92	0.88	0.45	1276
29	0.62	0.48	−0.51	−1.73	−0.87	0.93	0.88	0.44	1022
30	0.43	0.49	0.27	−1.93	0.09	1.03	1.03	0.37	452
31	0.70	0.49	0.35	−1.87	0.03	1.17	1.24	0.22	733
32	0.66	0.47	0.68	1.53	−1.13	0.99	0.99	0.37	1138
33	0.70	0.45	−0.91	−1.17	−1.30	0.82	0.74	0.53	1207
34	0.61	0.48	−0.46	−1.78	−0.85	0.92	0.89	0.45	990
35	0.53	0.49	−0.15	−1.97	−0.47	0.88	0.86	0.49	853
36	0.59	0.49	−0.40	−1.83	−0.70	0.98	0.99	0.38	840
37	0.82	0.37	−1.72	0.98	−2.02	0.94	0.90	0.35	1249
38	0.36	0.48	0.55	−1.69	0.30	0.93	0.91	0.45	480
39	0.29	0.45	0.89	−1.20	0.72	1.20	1.38	0.14	361
40	0.33	0.47	0.68	−1.54	0.48	0.97	1.01	0.41	384
41	0.46	0.49	0.13	−1.98	−0.20	1.12	1.17	0.25	791
42	0.38	0.38	0.38	0.38	0.32	1.05	1.06	0.34	405
43	0.36	0.48	0.55	−1.69	0.33	0.98	1.01	0.40	577
44	0.47	0.49	0.08	−1.99	−0.10	0.97	0.98	0.41	628
45	0.61	0.48	−0.49	−1.76	−0.77	0.88	0.84	0.48	776
46	0.32	0.46	0.75	−1.43	0.58	1.00	1.02	0.37	475
47	0.38	0.48	0.46	−1.78	0.35	0.99	1.01	0.40	445
48	0.59	0.49	−0.39	−1.84	−0.65	0.84	0.79	0.54	733

Table 18 (continued)

Descriptive statistics						*Rasch measurement*			
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
49	0.49	0.50	0.01	−2.00	−0.08	0.88	0.88	0.53	256
50	0.30	0.45	0.87	−1.24	0.76	0.91	0.92	0.47	281
51	0.25	0.43	1.12	−0.72	1.06	1.04	1.13	0.32	239
52	0.35	0.47	0.62	−1.60	0.52	1.08	1.13	0.28	461
53	0.49	0.50	0.02	2.00	−0.10	1.20	1.27	0.18	486
54	0.30	0.45	0.86	−1.25	0.83	1.13	1.18	0.27	233
55	0.52	0.49	−0.11	−1.99	−0.25	0.87	0.83	0.53	408
56	0.14	0.35	2.02	2.11	1.81	0.94	1.07	0.36	160
57	0.41	0.49	0.32	−1.89	0.23	1.18	1.22	0.24	360
58	0.42	0.49	0.31	−1.90	0.27	1.05	1.04	0.40	261
59	0.52	0.50	−0.08	−2.00	−0.22	0.98	0.97	0.45	297
60	0.45	0.49	0.18	−1.97	−0.12	1.23	1.30	0.21	308
61	0.50	0.50	−0.002	−2.00	−0.16	0.91	0.91	0.51	367
62	0.53	0.49	−0.15	−1.98	−0.28	0.88	0.85	0.52	483
63	0.51	0.50	−0.05	−2.00	−0.23	0.89	0.86	0.52	426
64	0.45	0.49	0.20	−1.96	0.23	1.15	1.21	0.29	297
65	0.34	0.47	0.65	−1.57	0.74	1.33	1.52	0.10	229
66	0.50	0.50	−0.018	−2.00	−0.09	1.03	1.00	0.44	268
67	0.63	0.48	−0.55	−1.70	-0.72	0.86	0.79	0.55	326
68	0.27	0.44	0.99	−1.01	1.17	0.98	1.06	0.41	173
69	0.61	0.48	−0.48	−1.76	−0.68	0.92	0.90	0.47	523
70	0.48	0.50	0.058	−2.00	−0.04	0.98	0.99	0.46	284

Table 19 Results of descriptive statistics analysis and Rasch measurement (IUUESEE 2018)

Descriptive statistics						*Rasch measurement*			
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
1	0.59	0.49	−0.37	−1.85	−0.76	0.92	0.89	0.47	1406
2	0.69	0.46	−0.82	−1.32	−1.32	0.91	0.87	0.46	2013
3	0.53	0.49	−0.15	−1.97	−0.56	0.88	0.84	0.52	1340
4	0.35	0.47	0.60	−1.63	0.37	0.98	0.98	0.41	930
5	0.21	0.41	1.35	−0.15	1.21	0.97	1.12	0.37	437
6	0.55	0.49	−0.20	−1.96	−0.42	0.91	0.88	0.51	664
7	0.37	0.48	0.49	−1.7	0.38	1.06	1.10	0.37	516
8	0.50	0.50	−0.02	−2.00	−0.36	0.98	0.96	0.42	1212
9	0.36	0.48	0.55	−1.69	0.38	0.96	0.99	0.44	709
10	0.52	0.49	−0.11	−1.98	−0.50	0.91	0.88	0.50	1126
11	0.91	0.28	−2.91	6.49	−3.10	0.94	0.67	0.36	3200
12	0.84	0.36	−1.90	1.62	−2.21	0.90	0.75	0.41	1966
13	0.77	0.41	−1.30	−0.30	−1.71	0.94	0.92	0.39	2089
14	0.67	0.46	−0.74	−1.44	−1.15	0.90	0.85	0.47	1510
15	0.54	0.49	−0.19	−1.96	−0.42	1.01	1.03	0.39	940
16	0.60	0.48	−0.44	−1.80	−0.69	0.94	0.93	0.45	904
17	0.79	0.40	−1.43	0.05	−1.82	0.86	0.72	0.47	2051
18	0.61	0.48	−0.46	−1.78	−0.75	0.95	0.95	0.42	1269
19	0.60	0.48	−0.42	−1.82	−0.82	0.96	0.92	0.43	1628
20	0.25	0.43	1.09	−0.79	0.99	1.07	1.16	0.32	438
21	0.16	0.37	1.76	1.11	1.61	0.94	1.17	0.38	197
22	0.38	0.48	0.49	−1.76	0.52	0.94	0.99	0.45	494
23	0.10	0.31	2.50	4.26	2.15	1.07	1.80	0.15	183
24	0.34	0.47	0.65	−1.56	0.47	0.98	1.03	0.41	707
25	0.18	0.38	1.66	0.76	1.48	0.92	1.05	0.39	399
26	0.51	0.50	−0.04	−2.00	−0.16	0.89	0.87	0.51	749
27	0.64	0.47	−0.59	−1.64	−0.79	1.18	1.79	0.19	627
28	0.46	0.49	0.12	−1.98	0.15	1.01	1.01	0.43	449
29	0.57	0.49	−0.30	−1.90	−0.45	0.89	0.87	0.51	689
30	0.40	0.49	0.38	−1.85	0.49	1.05	1.11	0.41	279
31	0.16	0.37	1.78	1.17	1.46	1.04	1.28	0.26	429
32	0.36	0.48	0.54	−1.70	0.30	1.00	1.02	0.40	720
33	0.30	0.46	0.83	−1.29	0.65	0.93	0.98	0.45	473
34	0.50	0.50	−0.03	−2.00	−0.46	1.30	1.45	0.07	1023
35	0.28	0.45	0.92	−1.14	0.65	1.07	1.06	0.32	476
36	0.49	0.50	0.03	−2.00	−0.24	0.97	0.95	0.43	835
37	0.62	0.48	−0.52	−1.72	−0.86	0.89	0.86	0.49	826
38	0.29	0.45	0.87	−1.23	0.57	1.30	1.50	0.11	294
39	0.24	0.43	1.16	−0.65	1.06	1.28	1.61	0.07	278
40	0.71	0.45	−0.92	−1.13	−1.27	0.95	0.95	0.40	1415
41	0.17	0.37	1.72	0.96	1.63	1.05	1.38	0.26	232
42	0.70	0.45	−0.90	−1.17	−1.19	0.96	1.18	0.36	1143
43	0.61	0.48	−0.47	−1.77	−0.64	0.85	0.80	0.53	755
44	0.40	0.49	0.38	−1.85	0.31	1.24	1.32	0.15	634
45	0.43	0.49	0.25	−1.93	0.18	1.15	1.19	0.28	556
46	0.44	0.49	0.21	−1.96	0.29	1.03	1.06	0.42	372
47	0.55	0.49	−0.21	−1.95	−0.43	1.04	1.09	0.36	569
48	0.27	0.44	1.03	−0.92	0.96	1.05	1.20	0.41	120

Table 19 (continued)

Descriptive statistics						*Rasch measurement*			
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
49	0.56	0.49	−0.27	−1.92	−0.33	0.93	0.90	0.49	432
50	0.16	0.37	1.79	1.21	1.76	1.13	1.96	0.16	182
51	0.33	0.47	0.69	−1.51	0.80	1.06	1.1	0.33	375
52	0.21	0.40	1.40	−0.01	1.54	1.20	1.61	0.17	183
53	0.62	0.48	−0.50	−1.74	−0.70	0.92	0.90	0.46	791
54	0.65	0.47	−0.63	−1.60	−0.77	0.88	0.85	0.51	490
55	0.54	0.49	−0.19	−1.96	−0.24	0.87	0.84	0.54	490
56	0.28	0.44	0.97	−1.04	0.98	0.97	0.96	0.43	280
57	0.47	0.49	0.096	−1.99	−0.06	1.11	1.16	0.32	610
58	0.56	0.49	−0.27	−1.92	−0.35	1.05	1.12	0.35	591
59	0.60	0.48	−0.44	−1.80	−0.68	0.99	0.99	0.39	994
60	0.50	0.50	−0.005	−2.00	−0.11	1.03	1.04	0.39	542
61	0.52	0.49	−0.12	−1.99	−0.48	1.19	1.41	0.28	301
62	0.36	0.48	0.55	−1.70	0.40	1.03	1.03	0.47	135
63	0.39	0.48	0.42	−1.82	0.33	1.02	1.06	0.45	215
64	0.50	0.50	−0.003	−2.00	−0.11	0.97	0.95	0.48	267
65	0.22	0.41	1.32	−0.24	1.31	1.16	1.46	0.33	88
66	0.38	0.48	0.45	−1.80	0.39	1.10	1.16	0.39	202
67	0.36	0.48	0.55	−1.69	0.53	1.12	1.20	0.32	283
68	0.35	0.47	0.59	−1.65	0.63	1.16	1.25	0.26	349
69	0.35	0.47	0.60	−1.63	0.46	0.89	0.87	0.52	384
70	0.57	0.49	−0.30	−1.91	−0.50	0.93	1.06	0.50	356

Table 20 Results of descriptive statistics analysis and Rasch measurement (IUUESEE 2019)

Descriptive statistics					*Rasch measurement*				
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
1	0.44	0.49	0.23	−1.94	−0.66	1.03	1.04	0.41	1323
2	0.64	0.47	−0.60	−1.64	−1.03	0.82	0.75	0.56	1585
3	0.22	0.42	1.28	−0.33	1.02	1.18	1.37	0.18	289
4	0.44	0.49	0.21	−1.95	−0.14	1.04	1.05	0.33	1208
5	0.79	0.40	−1.47	0.17	−1.88	1.15	1.58	0.09	2156
6	0.63	0.48	−0.55	−1.69	−0.99	0.93	0.89	0.44	1540
7	0.62	0.48	−0.49	−1.75	−0.80	0.91	0.88	0.45	1276
8	0.81	0.38	−1.61	0.62	−2.03	0.93	0.88	0.39	2037
9	0.37	0.48	0.52	−1.72	0.34	1.06	1.08	0.30	708
10	0.68	0.46	−0.81	−1.33	−1.34	1.05	1.11	0.29	1855
11	0.33	0.47	0.71	−1.49	0.43	0.94	0.99	0.42	725
12	0.82	0.38	−1.69	0.88	−2.02	0.87	0.73	0.45	1633
13	0.62	0.48	−0.50	−1.74	−0.88	0.87	0.83	0.50	1116
14	0.47	0.49	0.11	−1.98	−0.25	0.92	0.90	0.46	774
15	0.50	0.50	−0.01	−2.00	−0.33	0.90	0.89	0.47	1033
16	0.39	0.48	0.44	−1.80	0.16	1.15	1.20	0.21	776
17	0.49	0.50	0.001	−2.00	−0.19	1.03	1.08	0.35	651
18	0.20	0.40	1.46	0.14	1.31	1.08	1.38	0.23	218
19	0.20	0.40	1.46	0.14	1.27	1.19	1.38	0.09	426
20	0.75	0.43	−1.15	−0.65	−1.60	0.92	0.88	0.42	1857
21	0.11	0.32	2.39	3.75	2.10	1.01	1.32	0.23	168
22	0.31	0.46	0.80	−1.35	0.71	1.18	1.35	0.19	234
23	0.49	0.50	0.03	−2.00	−0.15	0.95	0.93	0.47	408
24	0.34	0.47	0.63	−1.60	0.18	1.19	1.31	0.29	148
25	0.33	0.47	0.68	−1.53	0.58	1.42	1.57	−0.02	347
26	0.79	0.40	−1.42	0.04	−1.69	0.88	0.75	0.45	1423
27	0.37	0.48	0.49	−1.75	0.41	0.97	0.98	0.39	570
28	0.53	0.49	−0.15	−1.97	−0.30	0.96	0.96	0.43	603
29	0.24	0.42	1.20	−0.55	1.17	0.89	0.94	0.44	359
30	0.19	0.39	1.54	0.38	1.48	1.05	1.31	0.23	250
31	0.41	0.49	0.34	−1.88	0.02	1.04	1.05	0.32	1030
32	0.59	0.49	−0.38	−1.85	−0.78	0.81	0.78	0.56	1037
33	0.14	0.34	2.07	2.30	1.66	0.99	1.20	0.26	292
34	0.38	0.48	0.47	−1.77	0.18	1.13	1.18	0.22	718
35	0.26	0.44	1.06	−0.86	0.90	1.04	1.13	0.29	428
36	0.56	0.49	−0.24	−1.94	−0.60	1.03	1.03	0.33	1001
37	0.45	0.49	0.19	−1.96	−0.12	0.93	0.91	0.44	828
38	0.04	0.21	4.18	15.5	2.98	1.09	2.50	−0.03	108
39	0.88	0.32	−2.39	3.71	−2.55	0.92	0.76	0.35	2106
40	0.62	0.48	−0.50	−1.74	−0.86	0.85	0.80	0.52	1430
41	0.31	0.46	0.77	−1.40	0.60	0.98	1.03	0.38	547
42	0.63	0.48	−0.57	−1.67	−0.79	0.86	0.79	0.51	920
43	0.31	0.46	0.80	−1.35	0.53	1.07	1.11	0.28	636
44	0.48	0.49	0.06	−1.99	−0.11	0.90	0.87	0.49	695
45	0.19	0.39	1.55	0.40	1.53	1.01	1.17	0.31	200
46	0.52	0.49	−0.08	−1.99	−0.30	1.01	1.03	0.36	830
47	0.45	0.49	0.18	−1.97	0.00	1.24	1.30	0.14	506
48	0.79	0.40	−1.49	0.24	−1.71	0.89	0.81	0.41	1181

Table 20 (continued)

Descriptive statistics					*Rasch measurement*				
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total scores*
49	0.59	0.49	−0.38	−1.85	−0.79	0.97	0.98	0.39	880
50	0.74	0.43	−1.10	−0.78	−1.25	0.90	0.83	0.46	717
51	0.71	0.45	−0.93	−1.12	−1.18	0.06	0.75	0.51	1058
52	0.42	0.49	0.31	−1.90	0.26	1.04	1.05	0.34	452
53	0.66	0.47	−0.71	−1.48	−0.90	1.00	1.09	0.34	729
54	0.30	0.30	0.30	0.30	0.83	1.17	1.38	0.19	275
55	0.22	0.41	1.34	−0.20	1.31	1.01	1.31	0.29	240
56	0.53	0.49	−0.15	−1.98	−0.33	0.88	0.85	0.50	747
57	0.35	0.47	0.62	−1.61	0.64	0.94	0.98	0.43	389
58	0.71	0.44	−0.97	−1.04	−1.13	0.91	0.87	0.44	829
59	0.54	0.49	−0.19	−1.96	−0.28	1.03	0.99	0.44	313
60	0.49	0.50	0.00	−2.00	−0.09	1.00	1.01	0.38	548
61	0.76	0.42	−1.25	−0.41	−1.62	0.86	0.75	0.48	1453
62	0.47	0.49	0.11	−1.99	0.03	1.01	1.04	0.45	255
63	0.24	0.42	1.19	−0.56	1.11	0.99	1.20	0.39	161
64	0.23	0.42	1.22	−0.50	1.14	1.19	1.40	0.22	165
65	0.30	0.46	0.84	−1.29	0.83	1.34	1.59	0.19	134
66	0.42	0.49	0.32	−1.90	0.15	0.89	0.91	0.52	345
67	0.34	0.47	0.66	−1.56	0.59	1.07	1.11	0.43	.59
68	0.07	0.27	3.10	7.65	2.63	1.12	2.81	−0.02	87
69	0.27	0.44	1.01	−0.96	0.98	0.97	1.05	0.42	184
70	0.25	0.43	1.15	−0.67	1.08	1.42	1.65	0.15	124

M mean, *SD* standard deviation, *MNSQ* mean square, *PT-measures* point-measure correlations

Acknowledgements
The authors wish to thank Professor Mike Linacre, Professor William Boone, and Dr. Vahid Aryadoust. We would not have been able to complete the analyses of the study without their expertise.

Authors' contributions
This manuscript was extracted from Hamidreza Babaee Bormanaki's dissertation. As a result, he handled the experiment, collected the raw data, and conducted data analyses under the supervision of his dissertation supervisor, Dr. Parviz Ajideh. The authors read and approved the final manuscript.

Authors' information
Hamidreza Babaee Bormanaki is a PhD candidate in the University of Tabriz in Iran. His research interests include test fairness, Rasch analysis, and DIF investigations. Dr. Parviz Ajideh is a professor of TEFL in the Department of English, Faculty of Persian Literature and Foreign languages, University of Tabriz. His research areas include language testing and ESP.

Funding
Not applicable.

Availability of data and materials
A copy of summarized descriptive and Rasch analyses can be provided as supplementary materials upon request. Individual participant responses to all test items cannot be provided at the present time because permission to publish or provide raw data (non-summarized) was not granted by the National Organization of Educational Testing.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 12 March 2022 Accepted: 12 August 2022
Published online: 10 September 2022

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.

Alavi, S. M., Karami, H., & Khodi, A. (2021). Test review of Iranian university entrance exam: English Konkur examination. *Language testing in Asia*, *11*(14), 1–10. https://doi.org/10.1186/s40468-021-00125-6.

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

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of IELTS listening test. *International Journal of Listening*, *8*(4), 40–60. https://doi.org/10.1080/10904018.2012.639649.

Aryadoust, V., Goh, C. M. C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, *8*(4), 361–385. https://doi.org/10.1080/15434303.2011.628632.

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing*, *33*(4), 529–553. https://doi.org/10.1177/0265532215594640.

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.

Belzak, W. C. M. (2019). Testing differential item functioning in small samples. *Multivariate Behavioral Research*, *55*(5), 722–747. https://doi.org/10.1080/00273171.2019.1671162.

Benıtez, I., Padilla, J. L., Hidalgo, M. D., & Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*, *29*(1), 1–16. https://doi.org/10.1080/08957347.2015.1102915.

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

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(Suppl 3), S176–S181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005.

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), 217–238. https://doi.org/10.1177/026553229901600205.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Sage.

Chalabi, M., & Janadele, A. (2007). The impact of cultural factors on economic achievement: A comparative study of Arab, Dezfuli and Lur Ethnicities in Khozestan Province. *Journal of Human Sciences*, *53*, 117–154.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*(2), 155–163. https://doi.org/10.1177/026553228500200204.

Chubbuck, K., Curley, W. E., & King, T. C. (2016). Who's on first? Gender differences in performance on the SATVR test on critical reading items with sports and science content. *ETS Research Report Series*, *16*(2), 1–116. https://doi.org/10.1002/ets2.12109.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. https://doi.org/10.1111/j.1745-3984.2005.00007.

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. 169–194). 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.

Golia, S. (2016). A proposal for categorizing the severity of non-uniform differential item functioning: The polytomous case. *Communications in Statistics - Theory and Methods*, *45*(2), 236–251. https://doi.org/10.1080/03610926.2013.830746.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). Confirmatory factor analysis of the Test of English as a Foreign Language. In *TOEFL Research Report*, (vol. 32, pp. 89–42). 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/10.1177/0265532211421161.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*, (pp. 97–111). Lawrence Erlbaum Associates.

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/10.1177/026553220101800104.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge: Cambridge University Press.

Kormos, J., & Kiddle, T. (2013). The role of socio-economic factors in motivation to learn to English as a foreign language: The case of Chile. *System*, *41*(2), 399–412. https://doi.org/10.1016/j.system.2013.03.006.
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), 225–252. https://doi.org/10.1177/026553229401100301.
Li, H., & Suen, H. K. (2012). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing*, *30*(2), 273–298. https://doi.org/10.1177/0265532212459031.
Lin, J., & Wu, F. (2003). Differential performance by gender in foreign language testing. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois.
Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266–283.
Linacre, J. M. (2010). *A user's guide to WINSTEPS*. Winsteps.com.
Linacre, J. M. (2012). *A user's guide to WINSTEPS*. Winsteps.com.
Linacre, J.M. (2021). Winsteps® Rasch measurement computer program (Version 5.1). Winsteps.com.
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/10.1177/0013164494054002003.
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. https://doi.org/10.1080/15434303.2011.565438.
Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286–300. https://doi.org/10.1080/08957347.2014.944305.
Oliveri, M. E., Lawless, R., Robin, F., & Bridgeman, B. (2018). 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. https://doi.org/10.1080/08957347.2017.1391258.
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. In *TOEFL Research Report*, (vol. 27, pp. 88–26). Educational Testing Service.
Pae, T. (2004). DIF for examinees with different academic backgrounds. *Language testing*, *21*(1), 53–73 Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.
Prieto Maranon, P., Barbero Garcia, M. I., Costas, S. L., & 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/10.1177/0013164497057004002.
Razmjo, S. A. (2006). Content analysis of specific questions of the English language test group of the national entrance exam of the country's universities. *Shiraz University Journal of Social Sciences and Humanities*, *1*(46), 465–480.
Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371. https://doi.org/10.1177/014662169602000404.
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). Sage.
Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*(1), 12–29. https://doi.org/10.1177/026553229200900103.
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/10.1177/026553229100800201.
Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In Holland, P. W. & Wainer, H. W. (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum, 281-315.
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.
Shin, J. Y. (2021). Investigating and optimizing score dependability of a local ITA speaking test across language groups: A generalizability theory approach. *Language testing*, *39*(2), 313–337. https://doi.org/10.1177/02655322211052680.
Shin, S. Y., Lee, S., & Lidster, R. (2021). Examining the effects of different English speech varieties on an L2 academic listening comprehension test at the item level. *Language testing*, *38*(4), 580–601. https://doi.org/10.1177/0265532220985432.
Şirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417.
Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*(3), 516–517.
Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3.
Suleman, Q., Hussain, I., Khan, F. U., & Nisa, Z. (2012). Effects of parental socioeconomic status on the academic achievement of secondary school students in Karak District. *International Journal of Human Resource Studies*, *2*(4), 14–31. https://doi.org/10.5296/ijhrs.v2i4.2511.
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). 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. In *TOEFL Research Report, 6*, (pp. 80–32). 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/10.1177/026553220001700303.
Timukova, A., & Drackert, A. (2019). What does the analysis of C-test gaps tell us about the construct of a C-test? A comparison of foreign and heritage language learners' performance. *Language Testing*, *37*(2), 235–253. https://doi.org/10.1177/0265532219861042.

Trace, J. (2019). Clozing the gap: How far do cloze items measure? *Language Testing*, *37*(2), 107–132. https://doi.org/10.1177/0265532219888617.
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/10.1191/0265532205lt301oa.
Vanbuel, M., & Deygers, B. (2021). Gauging the impact of literacy and educational background on receptive vocabulary test scores. *Language Testing*, *39*(2), 191–211. https://doi.org/10.1177/02655322211049097.
Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, *9*, 472.
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.
Xuelian, Z., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*, *35*(3), 412–436. https://doi.org/10.1080/09588221.2019.1704788.
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/10.1177/0013164402239316.
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. In *Educational testing Service*.
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.

Publisher's Note