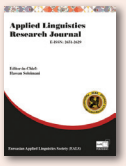
Applied Linguistics Research Journal

ALR Journal 2020;4(5):1–21

Original Research

doi: 10.14744/alrj.2020.73645

Investigating Gender-biased Items in a High-stakes Language Proficiency Test: Using the Rasch Model Measurement

Soodeh Bordbar,[1]
Seyyed Mohammad Alavi,[2]

[1]Assistant Professor at Department of English Language, Iran University of Medical Science, Tehran, *Iran*

[2]Professor at the Department of English and Foreign Languages, University of Tehran, Tehran, *Iran*

Corresponding Author: Soodeh Bordbar

Phone: +98-9126230337
e-mail: ut.sbordbar@yahoo.com

Article citation: Bordbar, S. & Alavi, S. M. (2020). Investigating gender-biased items in a high-stakes language proficiency test: Using the Rasch model measurement, *Applied Linguisics Research Journal, 4*(5): 1–21.

Received Date: April 19, 2020
Accepted Date: June 20, 2020
Online Date: September 5, 2020

Publisher: Kare Publishing

E-ISSN: 2651-2629

ABSTRACT

The consequential aspect of validity interprets the real and potential consequences of a test score, particularly when it comes to sources of invalidity related to the conceptions of fairness, bias, injustice, and inequity. Differential Item Functioning (DIF) analyzes the test items to evaluate test fairness and validity of educational tests. Besides, gender is mentioned as one of the elements that frequently acts as a source of construct-irrelevant variance. If gender imposes a large influence on the test items, it will bring about bias. In an attempt to explore validity and DIF analysis, the present study explores the validity of a high-stakes test and considers the role of gender as a source of bias in different subtests of language proficiency tests. To achieve this, the Rasch model was used to inspect biased items and to examine the construct-irrelevant factors. To obtain DIF analysis, the Rasch model was run to 5000 participants who were selected randomly from a pool of examinees taking part in the National University Entrance Exam for Foreign Languages (NUEEFL) as a university entrance requirement for English language studies (i.e., English literature, Teaching, and Translation). The findings reveal that the test scores are not free from construct-irrelevant variance and some misfit items were modified based on the fit statistics suggestions. By and large, the fairness of the NUEEFL was not confirmed. The results obtained from such psychometric assessment could be beneficial for test designers, stake-holders, administrators, as well as teachers. It also recommends the future administering standard and bias-free test and instructional materials.

Keywords: Differential Item Functioning analysis; Bias; Dimensionality; Fairness; The Rasch model.

1. Introduction

There has been increasing discussion about the importance of test use and consequences of tests in language teaching, learning, and assessment. The main concern in test development and test use is establishing the valid score-based interpretations and test scores use (Bachman, 1990). Test use is considered at the center of language assessment. As Bachman (1990, p. 55) suggests, "the single most important consideration in both the development of language tests and the interpretation of their results is the purpose or purposes which the particular tests are intended to serve". Moreover, he argues that "tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or

society at large" (Bachman, 1990, p. 279).

Looked at from a different perspective, questing and finding validity evidence becomes more salient as the stakes of the test becomes larger. To understand the importance of the consequences of the test may significantly differ across groups of stakeholders; "those who have to suffer the negative consequences, have a tendency to see them as more serious than those who do not suffer them" (Kane, 2013, p. 48).

Moreover, it is crucial for a test to be fair to different groups of test-takers. In other words, the test should not be inclined in favor of test-takers' characteristics, such as gender, ethnicity, academic background, etc. Gender is one of the factors that is frequently referred to as a source of construct-irrelevant variance. When the impact of the gender turns out to be significant, then the controversies over the tests' bias increases. And this issue definitely undermines the test validity. Besides, a statistical method needs to inspect such an issue to gauge whether the test items are differentially acting among groups of test-takers and, at least, investigate the source of construct-invariance. One of the recommended methods for achieving this goal is Differential Item Functioning (DIF) using the Rasch Model Measurement.

Further, every year a large number of test takers seat the National University Entrance Exam for Foreign Languages (NUEEFL) in Iran. One of the subjects in that exam is English. Besides, students are required to take several high-stakes tests to enter different university degrees. This paper focuses on finding evidence of the effects of gender performance among subtests of the language proficiency test. This is achieved by providing psychometric measurements of the validity of the proficiency test.

Present study addresses the following research questions:

1.Do the items of the test fit the Rasch model?

2.Is there any case of local item independence and unidimensionality among all the subtests of NUEEFL?

3.Is participants' gender a source of DIF in the subtests of the NUEEFL?

2. Literature review

2.1. High-stakes tests

High-stakes testing is one of the most provocative subjects in education, and the technicalities involved are highly complex. High-stakes and teacher-made test vary in the results' interpretations and consequences. In teacher-made test, scores and exams are interpreted in different ways. Failing in teacher-made tests could be interpreted as failure to learn the materials, whereas passing the test indicates mastery of the subject. Whilst The Glossary of Education Reform (2014) uses the scores obtained in high-stakes tests to assign punishments (such as sanctions, penalties, funding reductions, negative publicity), to receive compliments (awards, public celebration, positive publicity), to make advances (grade promotion or graduation for students), or to seek compensation (salary increases or bonuses for administrators and teachers). In the same vein, high-stakes tests aim to make "use of tests and assessments alone to make decisions that are of prominent educational, financial, or social impact" (Genesee & Upshur, 1996, p. 6). In addition, in high-stakes testing the goal is to provide all participants with an equal opportunity to determine their performance in the test (Song & He, 2015).

Moreover, the American Educational Research Association (AERA) published guidelines about high-stakes tests for policy makers with the intention of improving education and emphasizing meticulous evaluation in order to avoid the tests' potential to cause serious harm (Dunne, 2015).

High-stakes tests are also widely employed in Iran. Several studies have investigated various aspects of these tests (i.e., Farhady & Hedayati, 2009; Salehi & Yunus, 2012 a, b; Mirzaei, Hashemian, & Tanbakooei, 2012). Tahmasbi and Yamini (2012), for instance, have examined teachers' interpretations of students' scores and its impact on the future lives of the examinees taking the Iranian University Entrance Exam (IUEE). However, the results of factor analysis show that high-school teachers have no role in the IUEE development and administration processes. As highlighted

by teachers, neither language ability nor knowledge, played a part in the test. They believed that success or failure in high-stakes tests is largely explained by the use of test-taking skills, tactics, and strategies.

In another study, Sadeghi (2014) applied the Structuration Theory and Washback Hypothesis to investigate the effects of high-stakes testing in TOEFL and IELTS preparatory courses in Iran. He examined an interpretive ethnographic case study through observations and field notes to identify how high-stakes testing affected teachers' curriculum and methodology. On closer inspection, the teachers constantly encountered challenging questions which led to variations in their response to exam pressures.

Mirzaei et. al., (2012) pinpointed the transformative or reproductive practices of different stakeholders, including EFL teachers, learners, and parents with respect to the harmful impacts of the National University Entrance Test (NUET) in Iran. The results of the study revealed that EFL teachers' and high school students' actions and practices for the most part had a negative impact in teaching and practicing test-taking techniques before taking NUET. The harmful effects could also be seen when authentic materials were not used or the previous NUET items were reviewed in their class quizzes or final exams. A plethora of studies argued for and against making criteria in high-stakes testing which the stakes holders around the world are to follow (e.g., Qian & Cumming, 2017).

2.2. Construct validity

Validity has been considered as a unitary concept (Messick, 1980, 1989). The unified validity framework is a progressive matrix contains two facets, either evidence or consequence and either interpretation or use. Messick's framework suggests a number of research themes in the test validation focusing on the framework facets (Kunnan, 1998). Messick's interpretation of validity and his view of test consequences bring social impacts to the test's construct validity argument. As put forth by Mislevy, the Evidence-Centered Design framework is said to be inspired by Messick's validity matrix (Mislevy, Steinberg & Almond, 2002, 2003). Mislevy uses Messick's approach, however; he disregards the consequential facet of validity. Moreover, Mislevy concentrates on the test construction although Kane focuses on the test scores as a central part of the validity discussion (Kane, 1992, 2001).

Further, one main aspect of construct validity is the trustworthiness of score meaning and its interpretation. Construct validity, especially as it concerns high-stakes performance tests, involves the threats to the validity of the constructs. Further, the first threat to construct validity is construct under-representation relating to the ambiguity of score meaning and its interpretation. In other words, the tasks in the assessment exclude important facets of the construct. The test results are unlikely to display the test takers' true abilities within a construct. Thus, the results obtained limit the meanings of the score and its interpretation. In this regard, Messick (1989) stated that "the breadth of content specifications for a test should reflect the breadth of the construct invoked in score interpretation" (p. 35).

The second threat is construct-irrelevant variance meaning the test measures too many variables, which are irrelevant to the focal construct. These uncontrolled and extraneous variables affect the assessment outcomes. In fact, the inclusion of these sub-dimension variables is unavoidable. These variables show reliable variance in test scores, although they are irrelevant to the construct (Messick, 1989, 1996).

Regarding this, some sources of construct-irrelevant variance include poorly constructed examination questions, guessing, item bias, testing irregularities, and test-wiseness (Messick, 1989). As a matter of fact, construct-irrelevant variance leads to bias. Differential Item Functioning (DIF) analysis is one of the ways of investigating and eliminating bias items. DIF techniques are discussed below.

2.3. DIF analysis

Test developers use quality control or statistical procedures to ensure test items are proper and fair to all examinees (Camilli & Penfield, 1997; Holland & Wainer, 2012; Ramsey, 1993). These statistical procedures aim at identifying items with different statistical features across certain groups of examinees. This is referred to as differential item functioning (DIF) and "such items are said to function differentially across groups, which is a potential indicator of item bias" (Sireci & Rios, 2013, p. 170).

Wiberg (2007) stated that identifying problematic items via item analysis plays a key role in test assessment. He maintains that "item analysis includes using statistical techniques to examine the test takers' performance on the items" (Wiberg, 2007, p. 1). A crucial element in item analysis is detecting differential item functioning. The DIF technique is a very useful method for identifying potentially problematic items (Angoff, 1993). Moreover, DIF analysis can be employed to great advantage in second language (L2) assessments.

DIF occurs "when an item's properties in one group are different from the item's properties in another group" (Furr & Bacharach, 2007, p. 331). To highlight this point, Furr and Bacharach (2007) use the example of DIF in a situation where an item has different levels of difficulty for males and females.

DIF procedures are used to determine whether the individual item on a test function in the same way for two or more groups of examinees, "usually defined by their racial/ethnic background, sex, age/experience, or handicapping condition" (Scheuneman & Bleistein, 1989, pp. 255-256). There is a plethora of research studies that categorize DIF detection techniques. To date, many DIF analysis techniques have been proposed. With respect to DIF procedures, two major categories are presented by Scheuneman and Bleistein (1989, pp. 256-271) as:

> a). classical measurement methods. These methods make use of transformed item difficulty (TID), item discrimination procedures, partial correlation methods, contingency-table approaches, chi-square methods, log-linear methods, Mantel-Haenszel (MH) procedure, standardization, and distractor analysis
>
> b). item response theory (IRT)-based models which include three-parameter methods, and Rasch model.

Additionally, a wide range of possible techniques are available; however, only a limited number are currently used. For a comprehensive explanation of DIF detection methods, see the following: Kamata and Vaughn, 2004; McNamara and Roever, 2006; Scheuneman and Bleistein, 1989; Wiberg, 2007. The following section considers Item Response Theory (IRT)-based models, specifically the Rasch model as an applicable and germane method to present research.

2.4. The Rasch model

The Rasch model is mathematically equivalent to the one-parameter logistic (1PL) IRT model, but it developed separately (DeMars, 2010). However, there is controversy regarding the Rasch model. Some specialists believe that the Rasch and IRT models are structurally distinct and are used very differently. It is claimed that IRT models are used to "describe and fit data; when the fit is poor, the model is adapted or discarded in favor of another model" (Zand Scholten, 2011, p. 39). Conversely, the Rasch model is "more prescriptive. The data are required to fit the model and when they do not; items that show misfit are discarded until a satisfactory fit is obtained" (Zand Scholten, 2011, p. 39).

The Rasch model, just like the Pythagoras Theorem, is said to be a mathematical equation. It makes no predictions about item difficulties, in the same way that Pythagoras Theorem makes no predictions about the size of the triangle. The Rasch model and the Pythagoras Theorem both make predictions about the relationships between numbers. The Pythagoras Theorem posits a relationship among the three sides of the triangle. The Rasch model expresses the relationship between individual ability, item difficulty, and observed responses (Rasch Measurement Forum, 2017). Moreover, the concept of construct validity measures what it is intended to measure. For

instance, if we measure arithmetic ability, we expect the overall hierarchy of item difficulties to be addition, subtraction, multiplication, division. If this does not happen for an arithmetic test, then we cannot be certain what the test is measuring. And the test does not have construct validity (Rasch Measurement Forum, 2017).

It is crucial to keep in mind that investigating the similarities and differences among the Rasch model, IRT models conjoint measurement, Representational Measurement Theory (RMT), etc. is a different issue which is beyond the scope of this study. Also, these models require separate examination from the perspective of psychology, human science, statistics, and psychometrics. For a detailed discussion of the subject please refer to: Baker, 2001; Borsboom and Zand Scholten, 2008; DeMars, 2010; Michell, 1999, 2014; Ostini and Nering, 2006; Zand Scholten, 2011.

Put simply, the Rasch measurement helps to carefully design and modify a measurement instrument and calculate measurements that can be used with parametric statistical tests (Boone, Staver, & Yale, 2014). In other words, "psychometrics is the study of measuring psychological constructs and processes (knowledge, attitude, etc.) through the development of validation surveys, tests, and other assessments" (Boone et. al., 2014, p. 4). In fact, the employment of Rasch techniques indicates the use of psychometric analysis.

Furthermore, the three assumptions in the Racsh model are unidimensionality, local independence, and model fit. First, unidimensionality signifies one of the basic assumptions of the Rasch model. The precondition for this assumption is the logical notion that it is best to gauge one attribute at a time (Sick, 2010). A unidimensional test consists of items that refer to one dimension only. DeMars (2010) asserted that "whenever only a single score is reported for a test, there is an implicit assumption that the items share a common primary construct" (p. 38). He defined the unidimensionality as the model which specifies a single θ for each examinee, and any other factors have an impact on the item response which are treated as "random error or nuisance dimensions unique to that item and not shared by other items" and "violating this assumption may lead to mis-estimation of parameters or standard errors" (2010, p. 38). Sometimes, responses to the test items can be mathematically unidimensional, while the items measure what educators and psychologist would conceptualize as two different constructs. For example, test items may gauge test-taking speed and knowledge both (DeMars, 2010).

The second assumption is local independence. It states that "the probability of an individual responding correctly to a particular item is not dependent on previous responses or the responses given by other individuals to the same item" (Wale, 2013, p. 56). With tests of local independence, the spotlight is on dependencies among pairs of items. However, these dependencies are likely to appear as single dimensions.

The third assumption is data model fit estimate. There is constant interaction among persons, items, and the traits in Rasch analysis (Boone et. al., 2014). It is claimed that the Rasch model is "the only model that conforms to requirements of measurement", thus it is required to "take time to evaluate whether or not the data fit the model" (Boone et. al., 2014, p. 159). Put simply, Boone et al. (2014) explained that when we say a person or an item may be misfitting, what we mean is that an intended person and item do not act as the Rasch model predicts. The fit estimation checks for the model mis-specifications that evaluate the fit between the model and the data (DeMars, 2010). It is worth noting that if we have strong evidence that data from some participants sound odd, then their answers should not be used. Employing dubious data might discredit our analysis, so it is essential to first figure out whether the fit of the data to the model is satisfactory (Boone et al., 2014; Wale, 2013).

2.5. DIF, fairness, and bias

Although the significance of fairness has been ignored, it is presupposed that test developers and language researchers consider the concept of fairness when they investigate validity and reliability of the test. A number of researchers have recommended the use of DIF as an approach to address test fairness (e.g., Camilli & Shepard, 1994; Geranpayeh & Kunnan, 2007; Holland & Thayer,

1988; Kunnan, 1997, 2000, & 2004).

In fact the study of fairness in educational assessment goes back to at least the 1960s (Angoff, 1993). Kunnan (2010) asserts that test fairness has been one of the most fundamental concepts in evaluation entering the investigations of language assessment in the 1990s. As put forth by Xi (2010, p. 154) test fairness considers "comparable validity for all the identifiable and relevant groups across all stages of assessment, from assessment conceptualization to the use of assessment results". Therefore, ensuring fairness means absence of bias, which requires treating equally all test takers in testing procedure. Moreover, test fairness encompasses a wide area, such as embracing quality of management in test design, organizing and scoring, adequate construct validity, comprehensive and ample coverage of related content, items only measuring specific ability or skill without being influenced by construct-irrelevant factors, equal learning chances, and testing accessibility (Kunnan, 2000; McNamara & Roever, 2006; Saville, 2003; Shohamy, 2000).

3. Methodology

In following section, we will find information about the participants, instrumentations, the procedure of the study, and the process of data analysis.

3.1. Participants

The participants (N = 5000) were selected from among the pool of examinees from a population of 20,000 in Tehran, Iran, who had taken a recent version of the NUEEFL test. They were selected in a random manner from the two gender groups (i.e., males and females). Of the 5000 participants, 3335 were female and the remaining 1665 male. The academic background and the age of the participants were not an issue.

3.2. Instruments

The first instrument used in this research is the NUEEFL. It consists of a total of 95 items of which 25 are general English questions (from # 76 to 100). The other 70 items come under six subtests:

a.Grammar (10 items)-(from # 101 to 110)

b.Vocabulary (15 items)-(from # 111 to 125)

c.Sentence Structure (5 items)-(from #126 to 130)

d.Language Functions (10 items)-(from # 131 to 140)

e.Cloze Test (15 items)-(from #141 to 155)

f.Reading Comprehension (15 items)-(from # 156 to 170)

The NUEEFL test is annually administered to more than 100,000 applicants who want to get a bachelor's degree in foreign languages from a public university. All questions are in multiple-choice format and scored dichotomously. The test is time-restricted, lasting 105 minutes. In NUEEFL the correction for guessing is applied in the test items. In other words, guessing is not allowed with a total of three incorrect answers offsetting a correct answer.

Another instrument used for analyzing the data is Winsteps software (Version 3.92.1 updated in February 2016) (Linacre, 2016a, b). *Winsteps* constructs the Rasch measures using simple data sets (i.e., usually of persons and items) and applies the dichotomous Rasch model. This software can analyze combined item types, for instance dichotomous, multiple-choice, and multiple rating scales. It also examines in depth the structure of the items and persons. It uses a powerful diagnosis of multidimensionality via principle components analysis of residuals to detect and quantify substructures in the data (Linacre, 2016a). The version of *Winsteps* software (Version 3.92.1) used here can process up to 9,999,999 persons, 60,000 items, and up to 255 categories. It should be noted that *Winsteps* does not consider missing data or non-administered items.

Winsteps implements two methods of estimating the Rasch parameters, a) Joint Maximum Likelihood Estimation (JMLE), b) PROX (the Normal Approximation Algorithm devised by Cohen in 1979). In keeping with the methodology of the present study, *Winsteps* software implements the

JMLE method for estimating the Rasch parameters. In the JMLE formula, the estimate of the Rasch parameters happens when the observed raw score for the parameter matches the expected raw score. And the word "Joint" in JMLE means the simultaneous occurrence of the estimation of items and persons and rating scale structure of data matrix (adapted from *Winsteps* manual by Linacre, 2012; 2016a).

3.3. Procedure

This study focuses on the validity issue assessed through the application of the Rasch model. Given the nature of the study, the statistical and mathematical assumptions must be met. The steps followed are outlined below:

1.Preparing the data file for analysis, using SPSS and *Winsteps* software

2.Checking the data-model fit

3.Checking the assumptions of the Rasch model including, dimensionality and local independence

4.Analyzing DIF in the whole test and across the subtests

4. Results

Reliability analysis is a crucial part of any assessment. The *Winsteps* output provides the model requirements of an unbiased reliability estimate. The *Winsteps* tables provide a wide range of indices that can be used to evaluate the reliability of an instrument. The item separation index is from 0 to infinity, and the reliability index is from 0 to 1.

Reliability analysis is sensitive to the sample size which varies with the number of items and participants. As Linacre (2012) has said, the low item reliability means that "your sample is not big enough to precisely locate the items on the latent trait" (p. 644). The result of reliability analysis showed a reliability value of 1.00 which has high reliability for 95 test items. We can conclude with confidence that the high item reliability was also affected by the sample size of the data.

4.1. Checking the data-model fit estimate

The *Winsteps* software normally assesses the fit of the model through obtained statistics indicators of mean-square fit values (MNSQs) and the standardized Z values (ZSTDs). The values in the range of MNSQs are considered from zero to infinite (0- ∞) and the expected value is 1. Values above 1 likely show a deviation from the unidimensionality, and values less than 1 indicate an overfit in the response patterns with the data-model. The overfit in the model implies the existence of dependency among responses or items. Values between 0.70—1.3 are considered acceptable or so called good fit values and values less than 0.70 indicate overfit. Meanwhile, values above 1.3 signify underfit. Overfit for a low ability group indicates that the item is more discriminating between slight differences in ability.

In analyzing the model fit estimation, it is necessary to eliminate the participants with a total score of zero. The data were screened for outliers. As indicated earlier, the Rasch model does not estimate the zero scores and inevitably they are omitted from the analysis process. Hence, from a total of 5000 participants, 4965 remained for data analysis. Besides, the fit indices should be reported for the item calibration. The estimation of difficulty parameter and model fit estimations revealed that the range value of difficulty parameter is from 2.45 to -2.99, with a mean score of 0, and Standard Deviation (*SD*) of 1.21. The most difficult item is item Q.155 (Measure = 2.45) and the least difficult is the item Q.87 (Measure = -2.99).

It appears that 26 items were not located in the acceptable range. The range value of the outfit-MNSQs varies from 0.57 to 3.3 which denote that these items do not fit the model. The investigation of item statistics of outfit-MNSQs reveals that 26 items equal to 27% of items (i.e., 155, 126, 137, 105, 118, 166, 101, 103, 158, 111, 115, 133, 121, 167, 109, 128, 122, 156, 99, 84, 153, 149, 108, 160, 91, and 79) out of a total 95 items were not located in acceptable rating scale of 0.70 to 1.3. There

are four items from a total of 25 items in General Questions (i.e., 79, 84, 91, and 99), from total of 10 items, five items in Grammar (i.e., 101, 103, 105, 108, and 109), from total of 15 items, five items in Vocabulary (i.e., 111, 115, 118, 121, and 122), two items out of five items in Sentence Structure (i.e., 126, and 128), two items out of 10 items in Language Functions (i.e., 133, and 137), three out of 15 items in Cloze Test item (i.e., 149, 153, 155,), and from a total number of 15 items, five items in Reading Comprehension (i.e., 156, 158, 160, 166, and 167) did not fit the model.

The presence of a large number of misfitting items reveals that the data does not fit the model in each subtest. Therefore, the model and its assumptions may be violated. It is possible that the Rasch model unidimensionality assumptions also may or may not attain the desirable results. The following section reports the results of unidimensionality and local independence.

4.2. Checking unidimensionality and local independence

There are multiple methods for assessing dimensionality, including the data-model fit statistics. However, several studies have reported that these statistics do not have the ample sensitivity necessary for detecting multidimensionality. Besides checking the data-model fit, it makes sense to employ the Principal Components Analysis (PCA) on the raw data and residuals. In the present research, the Principal Components Analysis of residuals and a series of *t*-tests were used to check the unidimensionality of the test.

The following criteria were used to determine the unidimensionality of the test through the PCA analysis: a) if the amount of variance explained by measures be > 60%, b) "the unexplained variance of the eigenvalue for the first contrast (size) < 3.0 and unexplained variance explained by first contrast < 5% is good" (Linacre, 1991-2006, p. 272). Also, with regards to the criterion for eigenvalue, the expected eigenvalue is less than 2.0, but, in practice, a secondary dimension in the data usually requires a value of 3.0 or more.

In order to hold unidimensionality, there must be little residual correlation among items remaining. It is presumed that unidimensionality can be supported if 5% of *t*-tests are significant. If the level of significance of *t*-test is more than 5%, the local independency and unidimensionality will be violated.

The amount of the variance explained by different components in the data is 34.8% of which 12.7% is explained by persons and 22.1% by items. This indicated that a dominant first factor is present. As a rule of thumb, the variance explained by the first factor should be greater than 60% to be indicative of unidimensionality (Linacre, 2006). The result obtained here (34.8%, with an eigenvalue of 50.70) is lower than the minimum level necessary to demonstrate a unidimensional trait in the data. This showed that the items did not fit the model with item-person leveling. The first, second, third, fourth, and fifth unexplained variance with the eigenvalues of 3.4, 2.5, 2.2, 1.9, and 1.7 which were satisfactory according to the criteria. The results of the data analysis suggested that the unidimensionality does not hold across the whole test.

Local independence was examined through checking the ability parameter in order to identify whether the responses to items could be independent of each other (Pae, 2011). The Benjamini-Hochberg method was used to investigate the Rasch assumptions (Benjamini & Hochberg, 1995). First, the item difficulty parameter for all items was calculated using the Rasch model (it is called Level A). The investigation of outfit MNSQs statistics showed that 20 items had a value greater than 1.3. These 20 items confirm the unidimensionality assumption in data analysis. Second, the item difficulty parameters for these 20 items were determined using the Rasch model (it is called Level B).

The total sum of the differences between the difficulty parameter of these 20 items is -1.124 which is calculated from levels A and B. The constant correction value is -0.056 which is obtained by dividing -1.124 to the number of items (i.e., 20). As the next step, the ability parameter based on items in levels A and B was calculated. The results indicated that from a total of 5000 Student's *t*-statistics, 2680 or 53.6% were significant meaning that they were larger than the acceptable level of 5%. Thus, it is concluded that the unidimensionality and local independence assumptions do not hold in the entire test. To appreciate if unidimensionality and local independence hold in each

subtest, the results of item calibration in each of six subtests separately were analyzed.

In PCA, there are multiple ways to determine the factorability of inter-correlation matrix and to assess the appropriateness of using exploratory factor analysis. To determine the number of factors, some considerations such as Kaiser's criterion, Cattell scree-plot, and total variance explained should be taken into account. As shown in Table 1., for instance in Grammar section the Kaiser-Meyer-Olkin measure of sampling adequacy was 0.909 which is above the recommended value of 0.7. Also, the Bartlett's test of sphericity was significant (p- value = 0.00, $p < .05$).

Table 1.*KMO and Bartlett's Test in Grammar*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.909
Bartlett's Test of Sphericity	Approx. Chi-Square	9976.549
	df	45
	Sig.	0.000

The results revealed that all the extracted factors are not of interest to the researcher. The purpose of factor analysis is to explain the components with the smaller number of factors from the primary variable. First, it seeks to determine the number of factors or components that are kept in the factor analysis. In order to keep factors, usually mathematical criteria such as, Kaiser's criterion or Cattell Scree-plot test are employed. The Kaiser's cut-off value specifies the number of factors which have an eigenvalue of 1 or higher. Only those factors are kept which have a sum of squared factor loading (eigenvalues) equal or greater than 1. Some researchers keep enough factors to explain 80% of the variation. As shown in Table 2., the only factor with the value of 3.719 was component 1.

Table 2.*Total Variance Explained in Grammar*

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.719	37.193	37.193	3.719	37.193	37.193
2	.898	8.979	46.171			
3	.765	7.648	53.819			
4	.724	7.237	61.056			
5	.717	7.168	68.224			
6	.694	6.943	75.166			
7	.661	6.615	81.781			
8	.649	6.494	88.276			
9	.601	6.012	94.288			
10	.571	5.712	100.000			

The Cattell scree-plot displays the eigenvalues associated with a component or factor in a simple line plot in a diminishing size pattern. This scree-plot can be employed to graphically assess the optimal number of factors and to visualize factors which show most variability in the data. The ideal configuration in scree plot is a steep curve, followed by a bend and then a flat horizontal line. The place where the curve pattern for eigenvalues goes horizontal is called the scree point. The scree test shows the place where the smooth decrease of eigenvalues appears to level off to the right part of the plot. In the right side, the factorial scree is found. Put simply, the real operating factors are located on the left and the error operating factors are on the right (See Figure 1) (Ledesma, Valero-Mora, & Macbeth, 2015; Raîche, Walls, Magis, Riopel, & Blais, 2013).

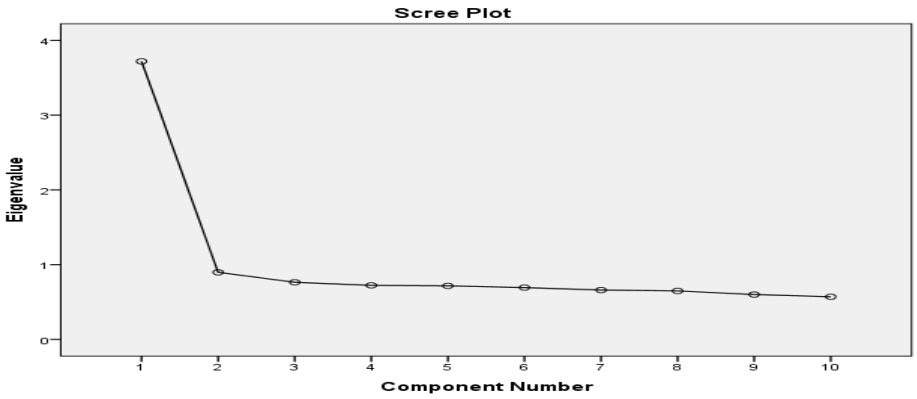

Figure 1. The scree-plot in Grammar

Regarding this, after selecting the components, we should perform factor matrix rotation. The main purpose of the rotation is to make the output more understandable by finding a simple structure. Rotations can be orthogonal (i.e., independence) and oblique (i.e., dependence/related). In a nutshell, the rotation was orthogonal and the results of component factor analysis displayed only one extracted component in a way that all variables were related to the first factor (See Table 3).

Table 3*Component Matrix in Grammar Subtest*

	Component
	1
X101	.534
X102	.640
X103	.648
X104	.600
X105	.643
X106	.554
X107	.588
X108	.652
X109	.608
X110	.620

Further, in each question of all subtests, extracting of the main factors and the PCA analysis were run. As a rule of thumb, a PCA analysis is significant if required variable for explaining 70% of variance is less than half of the variables. In the Grammar subtest the variance explained by the first factor was 37.193% is lower than the requirement of this criterion. It does not determine a unidimensional trait.

It can be inferred from the results of the PCA analysis in each subtest that the exploratory factor analysis was not significant for all six subtests. All in all, the results of the PCA analysis in the entire test were not significant. Besides, the unidimensionality of the trait in the data did not confirm. Moreover, there was no solid evidence of local independence across subtests and in entire test.

4.3. DIF analysis

Test developers use several quality controls or statistical procedures to make sure that the test items are appropriate for and fair to all examinees (Camilli & Penfield, 1997; Holland & Wainer, 2012; Ramsey, 1993). In this study, DIF analysis among all items across the entire test and subtests of the NUEFFL between male and female participants were investigated. When analyzing DIF in the Rasch

model, it is necessary to examine both the magnitude of the difference in logit units between groups and the statistical significance of the difference (Linacre, 2016a).

The magnitude of the DIF value should be at least 0.5 logits. In this phase of the study, DIF analysis was tested between groups of males and females. In order to examine the invariance, it is imperative to inspect the difference between the DIF analyses of these two groups by gauging the *t*-tests of the statistical significance of the data.

For statistically significant DIF, the probability of such differences (0.5 logits or larger) exists at random meaning ≤ 0.05. It is probable that such differences might crop up in the absence of systematic item bias among the test items (Linacre, 2016a).

The results obtained here revealed different difficulty levels in 85 items. They also showed that in 41 items the DIF contrasts were negative which means that these items are more difficult for female participants, whilst in 10 items the DIF contrast was zero. Further, in 44 items DIF contrasts were positive, demonstrating that these items were more difficult for the male group. The item difficulty has a normal distribution between gender groups.

Moreover, in 40 items the statistical differences between compared groups were significant. DIF Measure reports the difficulty (diff.) of each item for each person classification. In other words, DIF Measure is equal to DIF size plus the overall item difficulty (Linacre, 2016b). The difference in difficulty level showed that a large number of items were located above zero (See Figure 2). It denotes that the difficulty level of items in the NUEEFL was large. As proved by the DIF results, the NUEEFL test was a difficult test for the participants. And, the invariability of questions in gender groups was not accepted. Thus, the statement that the participants' gender is not a source of DIF in the NUEEFL is rejected.

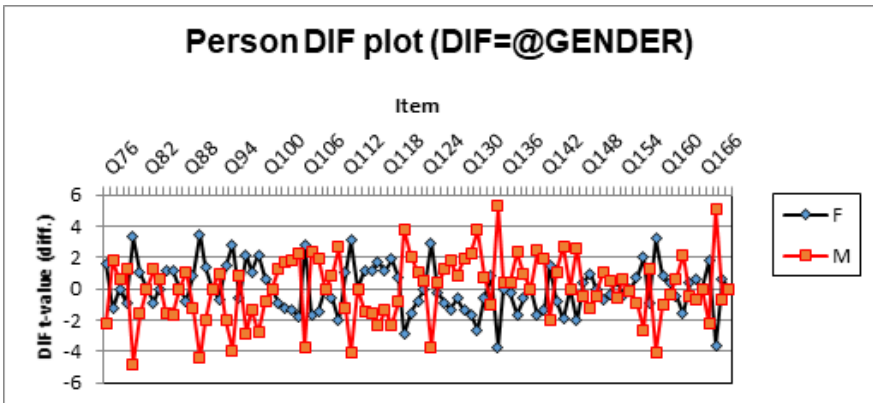

Figure 2. Comparing *t*-value difference between groups of males and females

However, an item that displays DIF is not essentially unfair to various groups of participants. And, it is too hasty to claim that the NUEEFL test is unfair to both male and female participants. In sum, finding the significance DIF between gender groups, the NUEEFL appeared not to be a DIF-free person estimates test. Hence, it is concluded that there is difference between males and females in answering the NUEEFL test. The following paragraphs present the results of the DIF analysis for the subtests. For instance, in the Reading Comprehension subtest the DIF analysis showed that 5 items out of 15 have significant DIF. As shown in Table 4, the highest variability of DIF contrast in the Reading Comprehension subtest is related to the item 168.

Table 4.*DIF results for Reading Comprehension Items*

Entry Number	Item	Sub Group	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch t	Rasch-Welch df	Rasch-Welch Prob
81	Q156	RC	M	0.44	F	0.54	-0.1	-1.12	INF	0.2624
82	Q157	RC	M	-0.04	F	0.24	-0.28	-3.3	INF	0.0010
83	Q158	RC	M	1.56	F	1.37	0.19	1.55	INF	0.1206
84	Q159	RC	M	0.47	F	0.97	-0.5	-5.12	INF	0.0000
85	Q160	RC	M	-0.28	F	-0.17	-0.1	-1.29	INF	0.1987
86	Q161	RC	M	0.62	F	0.66	-0.05	-0.5	INF	0.6143
87	Q162	RC	M	1.07	F	0.99	0.08	0.76	INF	0.4497
88	Q163	RC	M	0.05	F	-0.18	0.22	2.69	INF	0.0072
89	Q164	RC	M	0.2	F	0.26	-0.06	-0.66	INF	0.5119
90	Q165	RC	M	0.2	F	0.28	-0.08	-0.88	INF	0.3784
91	Q166	RC	M	1.57	F	1.57	0	0	INF	1.0000
92	Q167	RC	M	0.96	F	1.27	-0.3	-2.8	INF	0.0052
93	Q168	RC	M	-0.42	F	-0.89	0.47	6.22	INF	0.0000
94	Q169	RC	M	1.23	F	1.34	-0.1	-0.88	INF	0.3779
95	Q170	RC	M	1	F	1	0	0	INF	1.0000

Note1. Highlighted are the DIF flagged items in Reading Comprehension subtest.

Note2. RC = Reading Comprehension; M = Male; F = Female; INF = Infinity

Figure 3 displays the variance of ordering and spacing of all items. Except for two items (e.g., item 157 & 163), the remaining items have the same direction in ordering between compared groups of males and females. The results show that item 163 was an easy question in favor of the female group whereas item 157 is an easy question for male participants; it is interpreted as a male-favoring item. Except for items 157 and 163, the rest of DIF-Flagged items have the same ordering direction. And the difference of variability among items has been considered with the difficulty level of items in each subtest.

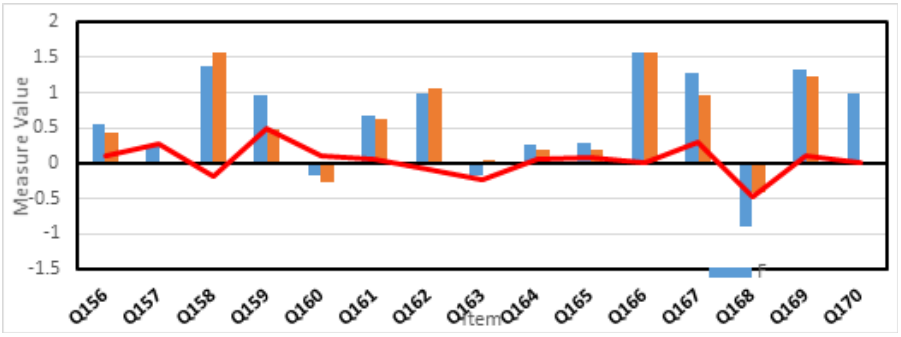

Figure 3. DIF size for reading comprehension items across gender groups

The results revealed that among 5 DIF-flagged items in the present subtest, the female group signified a lower degree of difficulty in 3 items. And, males indicated a lower degree of difficulty in the 2 other items. Thus, it is possible to conclude that the questions in the present subtest are easier for the female participants in comparison with the other group.

As discussed earlier, DIF analyses deliver a partial answer to fairness issues. Therefore, if grouping happens in a test and the items are favoring one group, then the test may not be fair enough for the other group. Hence, DIF analyses were investigated in the NUEEFL to resolve the matter. Among all subtests, General Questions, Sentence Structure, Language Functions, Cloze Test, and Reading

Comprehension turned out to be more female-favoring, whereas Vocabulary and Grammar were more favored by male participants (Please see Appendix for the DIF size figures in other Subtests). Our investigation has shown that a correct answer may require other knowledge, ability, and skill than the ones that the items aim to measure. Furthermore, the data analysis showed that these additional skills or knowledge are not equally presented between groups of males and females. And the items did not lead to the construct validity of the test for all the participants in gender groups because in some subtests the values obtained through DIF analysis were greater than expected (Uiterwijik & Vallen, 2005).

It has to be emphasized that four items (i.e., 108, 142, 157, & 163) were controversial in terms of the variance of direction in ordering. It can be inferred that these four items can be the source of bias. Moreover, an unequivocal result is that the favoring happens in the NUEEFL items and among subtests. Thus, the test is partial for the disfavored group.

5. Conclusions

There is an ongoing interest in comparing cultural, ethnic, or gender groups. DIF studies are absolutely essential in high- stakes testing programs. Furthermore, possible gender and/or ethnicity bias could negatively impact one or more groups as an irrelevant construct. In fact, the test administers attempt to develop perfectly fair test batteries; however, the dearth of research on the NUEEFL test makes it impossible to assess its fairness accurately.

The NUEEFL taken by tens of thousands of students annually acts as a gate-keeping test for those aspiring to enter the higher education system in Iran. In line with the main purpose of this research, DIF analysis was used to identify bias items across gender groups. It did not confirm a similar probability of endorsing the test items. The results of the study indicated that 40 items out of 95 turned out to be DIF-flagged items. This suggests that NUEEFL test scores are not free of construct-irrelevant variance. Hence, it does not support the argument for the construct validity.

To the best of the researcher's knowledge, the investigation of DIF analysis on the National Organization for Educational Testing's (NOET) data, the NUEEFL test, across gender groups was a brand new research project. This was particularly the case because the DIF analysis used the Rasch model in all sections of the national test. Whilst there are studies which investigate DIF across gender groups in language tests administered in Iran, they are principally concerned with the University of Tehran English Proficiency Test (UTEPT). Amirian, Alavi, and Fidalgo (2014), Rezaee and Shabani (2010), and Salehi and Tayebi (2011) are examples of such endeavors.

Our findings are in line with parts of the research performed on the UTEPT. In Amirian et.al. (2014) and Rezaee and Shabani (2010) DIF were displayed and observed in the different sections of the UTEPT. Although in some studies the results obtained were not compatible with our findings. For instance, Salehi and Tayebi (2011) have not found DIF items in reading section of the UTEPT. In another study Rayan and Bachman (1992) examined DIF with respect to the participants' performance in the TOEFL and the FCE exams. The results showed that the difference in the performance of males and females at the item level was negligible, whereas Carlton and Harris (1992) in a gender focused DIF study found that the females performed better than males. The results of the present research were compatible with Carlton and Harris's.

While that there are many gender focused DIF studies in a first language setting (i.e., Li & Suen, 2013; Ryan & Bachman, 1992; Zhang, Dorans, & Matthews-Lopez, 2003), few deal with DIF analysis in the context of English as a Foreign/Second Language (EFL/ESL) e.g., Alavi, Rezaee, and Amirian, 2011; Pae, 2004; Rezaee and Shabani, 2010; Salehi and Tayebi, 2012.

Rezaee and Shabani (2010) found significant DIF between gender groups in the UTEPT. The result of their study was verified by Karami (2011) who used the Rasch model to examine gender DIF in the UTEPT. The result of the study revealed that only 3 among 19 DIF items displayed practical significant DIF. Also, Amirian et al., (2014) detected gender DIF with the UTEPT. They performed a twofold research. The result of the first phase revealed that there is substantial DIF between gender groups in UTEPT. In the second phase, they performed content analysis on the DIF-flagged items to understand the source of DIF; and the result showed that humanities-oriented topics were mainly

female favoring, while science-oriented topics were mostly favored by males. In fact, the literature on gender DIF in EFL/ ESL context using the Rasch model is inadequate (Karami, 2010, 2015).

Moreover, several studies have investigated DIF at the language skills level. For instance, Aryadoust, Goh, and Kim (2011) examined gender DIF in the Michigan English Language Assessment Battery (MELAB) listening section using the Rasch measurement. The result of the uniform DIF (UDIF) revealed that two items favor different gender groups. Also, the non-uniform DIF (NUDIF) analysis showed several items with significant DIF mostly favoring low proficient male participants.

To bridge this gap, the present research attempted to validate the NUEEFL test in the case of gender. The present gender DIF study implements the Rasch model to figure out whether the NUEEFL as a high-stakes test shows substantial DIF in favor of a specific gender group. The results indicated that the NUEEFL test was favored by females. Apparently, disfavored group (i.e., male group) was not treated fairly. The test turned out to be unfair.

The results of the present study are controversial due to the statement of Rezai-Rashti and Moghadam (2011). They believed that a range of restrictions have been put on the number of female students that can enroll in each field to restrict females altogether from certain majors; whereas the result of the present research confirmed that the test is administered in favor of females.

With respect to research conducted by foreign scholars, the work of Lin and Wu (2003) is very similar. They employed the computer program CIPTEST for DIF/DBF (Differential Bundle Functioning) analysis and examined dimensionality of the English Proficiency Test in China. Their work shows much greater gender DIF in the overall test and among subtests. In another study, Tae (2004) conducted DIF analysis in the Reading Comprehension section of Korean National Entrance Exam for Colleges and Universities whose results showed gender DIF in Reading Comprehension section of the exam.

Additionally, the results of the present study are consistent with Karami's (2015) with respect to dimensionality of NUEEFL, in which the multidimensionality in the whole test and among sub-tests was proven to exist. The results of the present research showed that the NUEEFL is not unidimensional.

Fairness and DIF analysis are broad concepts which involve analyzing the DIF among the items, considering the performance of different groups, or focusing the bias on the performance of every individual participant (See Camilli, 2006; Xi, 2010; Kunnan, 2010). Test bias is related to the issue of test fairness, and the social ramifications of test results have unfairly advantaged or disadvantaged specific groups of test takers.

The university entrance exam often raises concerns about the issues of test fairness and bias. Since Messick (2013) has noted that validity and fairness are a matter of degree. As Messick (2013) has pointed out, a test is neither absolutely valid nor absolutely invalid. Also, the fairness of the test is not absolute. A test can be to some degree fair or unfair. The results of the present study show that the DIF were significant among 40 items. Among subtests of NUEEFL (i.e., Sentence Structure, Language Functions, Cloze Test, and Reading Comprehension) turned out to be more female-favoring. Whereas Vocabulary and Grammar were more favored by male. And among DIF items, four items (i.e., 108, 142, 157, & 163) had critical results regarding the variance of direction in ordering. And the construct validity of the NUEEFL was threatened in compared groups.

It should be noted that in present research only receptive skills including reading comprehension, vocabulary, and grammar, were examined. Other skills such as listening, writing, and speaking were not included. In listening comprehension, for instance, most females were found to have an advantage compared to males (Boyle, 1987; Cole, 1997). The results of analysis indicated significant DIF between males and females and the findings from this study are not consistent with the results of Ryne and Bachman's (1992) who found no gender difference in any of TOEFL subtests.

6. Implications of the study

The major contribution of this study is to the field of language testing by providing empirical evidence for the interpretation of the NUEEFL items and scores via a comprehensive investigation of the item fit, dimensionality, and the detection of biased items. As Cole and Zieky (2001) have pointed out, the fairness issue has been the priority in educational assessments.

As a standard procedure for any large scale assessment, DIF evaluation is proposed by American Psychological Association (APA), and American Educational Research Association (AERA) (Huang & Han, 2012). DIF is a necessary but not sufficient condition for the test or item bias. Abbott (2007) elaborated on the implementation of DIF analysis. It is maintained that the statistical methods are useful for flagging DIF items. As suggested, content analysis is essential to appreciating why the items function differentially between the groups and to understand the nature of DIF performing.

According to Camilli (2006) DIF analysis mainly focuses on the performance of two or more different groups. Therefore, such analysis cannot disclose the existence of bias against different individuals. Moreover, this study draws attention to the fact that the evaluation of the high-stakes tests, such as the NUEEFL, requires us to pay attention to individual performance rather than group performance. In this regard, it is proposed that the students should be clustered based on their weak points as detected through the University Entrance Exam. Then, after admission to university, the students' weaknesses in any of the subtests should be reported to the relevant authorities. After that universities will be able to offer students English courses to help remedy their weaknesses.

Moreover, this study makes some recommendations for test designers, stake-holders, administrators, as well as teachers and students. By and large, it can be beneficial for anyone who is associated with high-stakes tests. The current research provides a good deal of information about DIF analysis in high-stakes test by application of the Rasch model approach. It examines gender groups and their performance on the NUEFL test. The Rasch model determines the characteristics of items individually (i.e., item difficulty and item discrimination) as well as identifies the participants' ability as matched to the certain items.

Additionally, the results of this study are directed towards teachers and decision makers, especially those in the National Organization for Educational Testing (NOET). It has solid implications for educational policy and practice. Furthermore, evaluating constructs and test fairness for the university entrance exam, and specifically the NUEEFL is essential because it currently is used to gauge high school students' knowledge by the National Organization for Educational Testing (NOET).

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, 24(1), 7-36.

https://doi.org/10.1177/0265532207071510

Alavi, S. M., Rezaee, A. & Amirian, S. M. R. (2011). Academic discipline DIF in an English language proficiency test. *Journal of English Language Teaching and Learning, 5(7),* 39-65.

Amirian, S. M. R., Alavi, S. M., Fidalgo, A. M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, Vol. 4, No. 2.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Erlbaum.

Aryadoust, V., Goh, C. C. M. & Kim, L. O. (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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Baker, F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1., pp. 289-300, doi:10.2307/2346101.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences.* Dordrecht: Springer Science, and Business Media.

https://doi.org/10.1007/978-94-007-6857-4

Borsboom, D., & Zand Scholten, A. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory Psychology, 18,*111-117.

https://doi.org/10.1177/0959354307086925

Boyle, J. (1987). Sex differences in listening vocabulary. *Language Learning, 37*(2), 273-

284.

https://doi.org/10.1111/j.1467-1770.1987.tb00568.x

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., Vol. 4, pp. 221-256). Westport, CT: American Council on Education & Praeger.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on the Mantel-Haenszel log-odds ratio. *Journal of Educational Measurement, 34,* 123–139.

https://doi.org/10.1111/j.1745-3984.1997.tb00510.x

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

Carlton, S. T., & Harris, A. M. (1992).*Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons.* Princeton, NJ: Educational Testing Service.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *The British Journal of Mathematical and Statistical Psychology, 32,* 113-120.

https://doi.org/10.1111/j.2044-8317.1979.tb00756.x

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

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01132.x.

https://doi.org/10.1111/j.1745-3984.2001.tb01132.x

DeMars, C. (2010). *Item Response Theory: Understanding statistics measurement*. Oxford, UK: Oxford University Press.

https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Dunne, D. W. (2015). Cautions issued about high-stakes tests. *Education World: http://www.educationworld.com/a_issues/issues110.shtml*

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics, 29,* 132–141.

https://doi.org/10.1017/S0267190509090114

Furr, M. R., & Bacharach, V. R. (2007). *Psychometrics: An introduction*. Thousand Oaks, CA: SAGE.

Glossary of education reform for journalists, parents, and community members (2014). *http://edglossary.org/high-stakes-testing/School Communications Contact*.

Genesee, F., & Upshur, J. (1996). *Classroom-based evaluation in second language education*. Cambridge University Press.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential Item Functioning in terms of age in the Certificate in Advanced English Examination. *Language Assessment Quarterly, 4, 190-222.*

https://doi.org/10.1080/15434300701375758

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

Holland, P. W., & Wainer, H. E. (2012). *Differential item functioning*. London, UK: Routledge.

https://doi.org/10.4324/9780203357811

Huang, j., & Han, T. (2012). Revisiting Differential Item Functioning: Implications for Fairness Investigation. *International Journal of Education*. Vol. 4, No.2, 74-86.

https://doi.org/10.5296/ije.v4i2.1654

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. Learning Disabilities: *A Contemporary Journal*, 2, 49-69.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

https://doi.org/10.1111/jedm.12000

Karami, H. (2010). A differential item functioning analysis of a language proficiency test: an investigation of background knowledge bias. Unpublished MA Thesis, University of Tehran.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5(2),* 27-38.

Karami, H. (2015). A closer look at the validity of the University Entrance Exam: Dimensionality and generalizability. (Unpublished Ph.D dissertation, University of Tehran).

Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta, V. Kohonen, L. Kurki-Suomo, & S. Luona (Eds.), *Current developments and alternatives in language assessment* (pp. 85–105). Jyväylä, Finland: University of Jyväylä.

Kunnan, A. J. (1998). Validation in Language Assessment: selected papers from the 17th Language Testing Research Colloquium, Long Beach. *Lawrence Erlbaum Associates. Inc. Printed in USA.* (pp. 1-275).

Kunnan, A. J. (2000). *Fairness and justice for all. In A. J. Kunnan (Ed.), Fairness and validation in language assessment (pp. 1–14)*. Cambridge: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic and C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge: Cambridge University Press.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183–189.

https://doi.org/10.1177/0265532209349468

Ledesma, R. D., Valero-Mora, P., & Macbeth, G. (2015). The scree test and the number of factors: a dynamic graphics approach. *The Spanish journal of psychology*, *18*, E11.

https://doi.org/10.1017/sjp.2015.13

Li, H., & Suen, H. (2013). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing. 30,* 273-298. 10.1177/0265532212459031.

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, IL.).*

Linacre, J. M. (1991-2006). A user's guide to Winsteps® Ministep Rasch-model computer programs. Retrieved

January, 10, 2007, from http://www.winsteps.com/aftp/winsteps.pdf

Linacre, J. M. (2006). Data variance explained by measures. *Rasch Measurement Transactions, 20,* 1045–1047.

Linacre, J. M. (2012). A user's guide to Winsteps [User's manual and software]. Retrieved from http://www.winsteps.com/winsteps.htm.

Linacre, J. M. (2016a). Winsteps® Rasch measurement computer program User's Guide. *Beaverton, Oregon:* Retrieved from http://www.winsteps.com/.

Linacre, J. M. (2016b). Winsteps® (Version 3.92.1) [Computer Software]. *Beaverton, OR: Winsteps.com.* Retrieved from http://www.winsteps.com/.

McNamara, T., & Roever, C. (2006) *Language testing: The social dimension*. Malden, MA: Blackwell.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.

https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*. (3rd ed.) (pp. 13–103). New York: American Council on Education & Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing 13 (3),* 241-256.

https://doi.org/10.1177/026553229601300302

Messick, S. J. (Ed.). (2013). *Assessment in higher education: Issues of access, quality, student development, and public policy.* Routledge, Taylor and Francis Group.

https://doi.org/10.4324/9781315045009

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* New York: Cambridge University Press.

https://doi.org/10.1017/CBO9780511490040

Mirzaei, A., Hashemian,M., & Tanbakooei, N. (2012). Do Different Stakeholders' Actions Transform or Perpetuate Deleterious High-Stakes Testing Impacts in Iran?. *The 1st Conference on Language Learning & Teaching: An Interdisciplinary Approach (LLT –IA).*

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496.

https://doi.org/10.1191/0265532202lt241oa

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of assessment arguments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models. Quantitative applications in the social sciences.* Thousand Oaks, CA: SAGE.

https://doi.org/10.4135/9781412985413

Pae, T. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32(2),* 265–281.

https://doi.org/10.1016/j.system.2003.09.009

Pae, H. (2011). Differential item functioning and unidimensionality in the Pearson Test of English Academic. *Pearson Education Ltd.*

Qian, DD, & Cumming, A. (2017). Researching English language assessment in China: Focusing on high-stakes testing. *Language Assessment Quarterly*, *14*(2), 97–100.

https://doi.org/10.1080/15434303.2017.1295969

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Erlbaum.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J. G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 23.

https://doi.org/10.1027/1614-2241/a000051

Rezaee, A. A., & Shabani, E. (2010). Gender differential item functioning analysis of the University of Tehran English Proficiency Test. *Pazhuhesh-e Zabanha-ye Khareji.* No. 56. pp. 89-108.

Rezai-Rashti, G., & Moghadam, V. (2011). Women and higher education in Iran: What are the implications for employment and the "marriage market"? *International Review of Education, 57,* 419–441.

https://doi.org/10.1007/s978-94-007-4411-0_10

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

Sadeghi, S. (2014). High-stake Test Preparation Courses: Washback in accountability contexts. *Journal of Education & Human Development*. March. Vol. 3, No. 1, pp. 17-26.

Salehi, M. & Tayebi, A. (2012). Differential item functioning in terms of gender in reading comprehension subtest of a high-stakes test. *Iranian Journal of Applied Language Studies*. Vol. 4, No. 1. pp. 135- 168.

Salehi, H. & Yunus, M.M., (2012a). The Washback Effect of the Iranian Universities Entrance Exam: Teachers' Insights. *GEMA Online™ Journal of Language Studies*. Volume 12(2), pp. 609- 628.

Salehi, H. & Yunus, M.M., (2012b). University Entrance Exam in Iran: A bridge or a dam. *Journal of Applied Sciences Research,* 8(2): 1005-1008, 2012. ISSN 1819-544X

Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English examination1913–2002* (pp. 57–120). Cambridge: Cambridge University Press.

Scheuneman, J. D., & Bleistein, C. A. (1989) A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education.* 2, 255-275.

https://doi.org/10.1207/s15324818ame0203_6

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

Sick, J. (2010). Assumptions and requirements of the Rasch measurement. *JALT Testing & Evaluation SIG Newsletter, 14*(2), 23-29.

Sireci, S. G., & Rios, J. A., (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19, (2–30), 170–187, doi.org/10.1080/13803611.2013.767621.

https://doi.org/10.1080/13803611.2013.767621

Song, X. & He, L. (2015). The effect of a national education policy on language test performance: a fairness perspective. *Language Testing in Asia*, *5:4.* DOI 10.1186/s40468-014-0011-z.

Tae, P. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, *32,* 265-281.

https://doi.org/10.1016/j.system.2003.09.009

Tahmasbi, S., & Yamini, M.. (2012). Teachers' Interpretations and Power in a High-Stakes Test: A CLA Perspective. *English Linguistics Research, 1*(2), p53. doi: 10.5430/elr.v1n2p53.

https://doi.org/10.5430/elr.v1n2p53

Wale, C. M. (2013). Evaluation of the effect of a digital mathematics game on academic achievement. (Doctoral dissertation, University of Northern Colorado).

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods. *Educational Measurement,* technical report N. 2.

Xi, X. (2010) How do we go about investigating test fairness? *Language Testing,* 27(2), 147-170.

Zand Scholten, A. (2011). Admissible statistics from a latent variable perspective. *The Institutional Repository of the University of Amsterdam (UvA),* 29-46.

Zhang, Y., Dorans, N. J. & Matthews-Lopez, J. L.(2003) .Using DIF dissection method to assess effects of item deletion. Research Report No. 2005-10. College Board.

https://doi.org/10.1002/j.2333-8504.2005.tb02000.x

Rasch measurement forum. (2017). Retrieved from http://raschforum.boards.net/.

Appendix

The DIF size Figures in Other Subtests

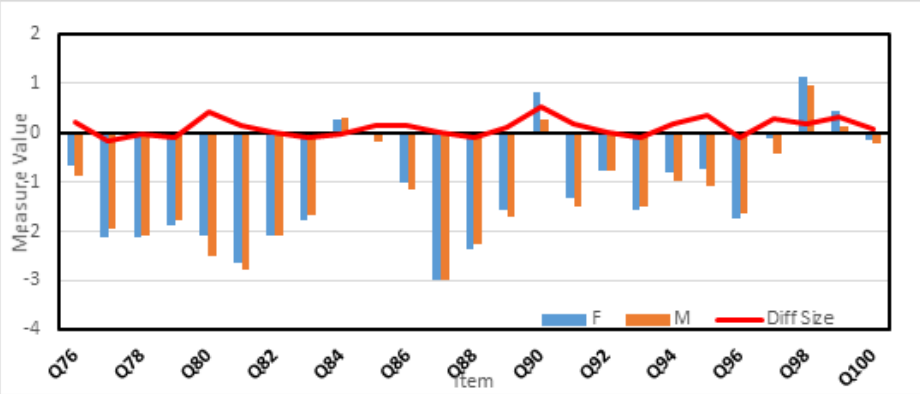

Figure a. DIF size for general questions across gender groups

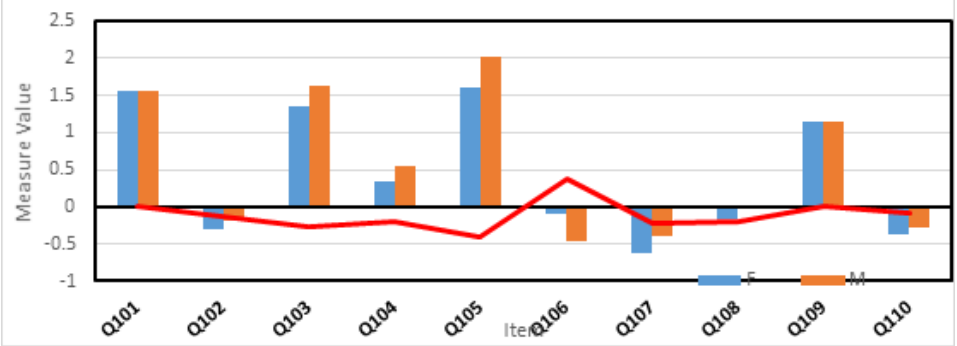

Figure b. DIF size for grammar items across gender groups

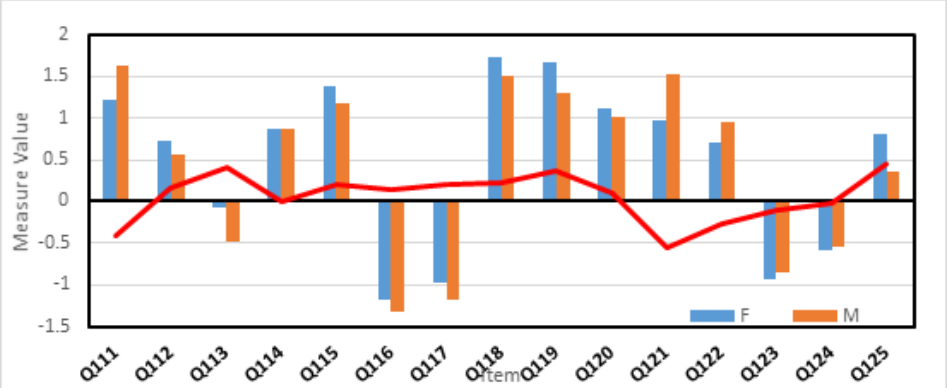

Figure c. DIF size for vocabulary items across gender groups

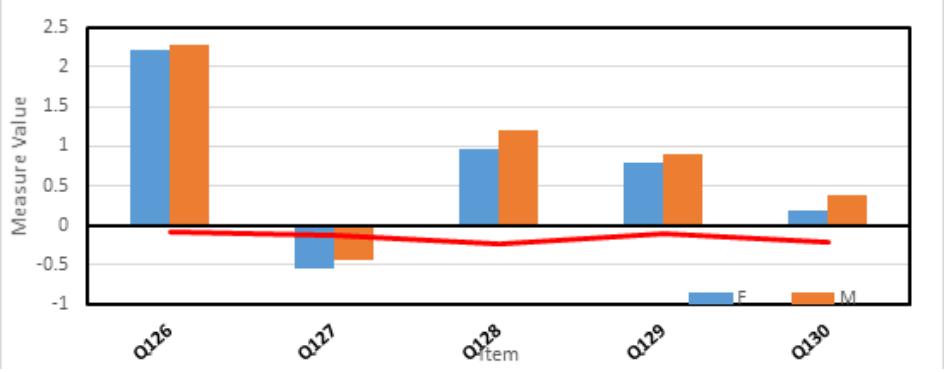

Figure d. DIF size for sentence structure items across gender groups

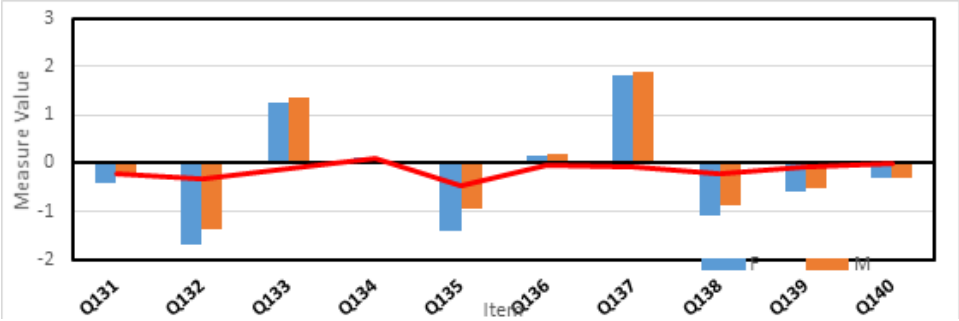

Figure e. DIF size for language function items across gender groups

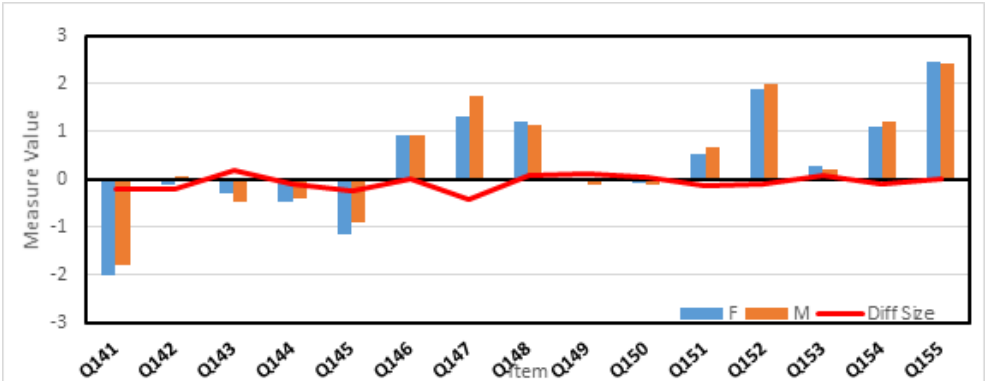

Figure f. DIF size for cloze test items across gender groups