Differential Item Functioning Analysis of High-Stakes Test in Terms of Gender: A Rasch Model Approach

[1] University of Tehran, Tehran, Iran
Email:Mohammed.Alavi@gmail.com

[2] PhD candidate of TEFL, Tehran University, Tehran, Iran
Email: ut.sbordbar@yahoo.com

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

ABSTRACT

Differential Item Functioning (DIF) analysis is a key element in evaluating educational test fairness and validity. One of the frequently cited sources of construct-irrelevant variance is gender which has an important role in the university entrance exam; therefore, it causes bias and consequently undermines test validity. The present study aims at investigating the presence of DIF in terms of gender in a high stakes language proficiency test in Iran, the National University Entrance Exam for Foreign Languages (NUEEFL). The participants' responses (N = 5000) were selected randomly from a pool of examinees who had taken the NUEEFL in 2015. The results displayed DIF between male and female test takers. Hence, on the basis of the findings, it is concluded that the NUEEFL test scores are not free of construct-irrelevant variance and the overall fairness of the test is not confirmed. Also, both Rasch assumptions (i.e., unidimensionality and local independence) are hold in the present research.

Keywords: *Differential Item Functioning, Dimensionality, Rasch Model*

INTRODUCTION

In language testing and educational measurement the discussions about test use and the consequences of tests have increased. Since the National University Entrance Exam for Foreign Languages (NUEEFL) is administered annually to a large number of test takers country-wide in Iran, the consequences of failure on the test are serious. It could result in spending one or more years for test preparation and two-year military service (for males).

Therefore, it is essential to examine the extent to which the instrument assesses what it is intended to measure (validity) as well as the test consistency (reliability) (Pae, 2011) in measuring the English ability in the high-stakes test, such as NUEEFL. Nonetheless, despite the heated nature of the debates, there has been little empirical evidence for the validity of the NUEEFL test and its fairness. Specifically, there is no ample evidence of test fairness among male and female test takers. In the absence of such evidence, any talk of the fairness of the selection policy would be doomed to fail.

The present study aims at investigating the validity of a high-stakes test in general and to considering the role of gender as a source of bias in the NUEEFL, in particular. Regardless of the content of the debates over the gender issue, it appears that there is no evidence on the effect of gender on the performance on the NUEEFL. If gender asserts a large influence, then it would be a case of bias and will undermine validity of the test. This is because gender is not part of the construct measured by the test and any significant impact by gender is a case of construct-irrelevant variance. As a part of standard process, Differential Item Functioning (DIF) analysis is conducted on the test items, as a main factor in the evaluation of the fairness, and validity of educational tests.

In order to investigate the psychometric properties of the high-stakes test (i.e., NUEEFL), the present study will address the following research questions:

1. To what extent do the item responses of NUEEFL form a unidimensional construct according to the Rasch measurement model?
2. Is participants' gender a source of DIF in NUEEFL items?

Review of Related Literature

Differential Item Functioning (DIF)

Test developers deploy several quality control or statistical procedures to ensure that the test items are proper and fair for all examinees (Camilli & Penfield, 1997; Holland & Wainer, 2012; Ramsey, 1993). The statistical procedure aims at identifying items with different statistical features across certain groups of examinees. This refers to 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).

According to Geranpayeh and Kunnan (2007) irrespective of considering the fairness issue in the design-development-administration-scoring cycle, still many problems are found in this procedure. Geranpayeh and Kunnan (2007) maintained there are two approaches for solving these problems. One approach is to develop a pilot group in order to examine test scores. If the test has been already conducted, a large sample is available to examine test scores and to investigate items and functions. If it is identified that they act differently, the source of this difference is known as Differential Item Functioning (DIF).

To characterize the definition of DIF, Wiberg (2007) indicated that identifying problematic items via item analysis plays a key role in a test. It is maintained that "item analysis includes using statistical techniques to examine the test takers' performance on the items" (Wiberg, 2007, p. 1) and one of the crucial parts in the item analysis is to detect differential item functioning. The DIF technique is still a very useful method for identifying potential problem items (Angoff, 1993).

Differential item functioning 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) specified via an example; DIF exists when a particular item has different levels of difficulty for males and females. Put another way, the incidence of differential item functioning means that a male and a female who have the same trait or ability level have different probabilities of answering the item correctly. It is concluded that the presence of DIF between groups shows that the groups cannot be meaningfully compared on the item (Furr & Bacharach, 2007).

DIF procedures are used to determine whether the individual items on a test function in the same way for two or more groups of examinees, "usually defined by racial/ethnic background, sex, age/experience, or handicapping condition" (Scheuneman & Bleistein, 1989, pp. 255-256). A plethora of studies categorized DIF detection techniques. To date, many DIF analysis techniques have been proposed. McNamara and Roever (2006, p. 93) classified methods for detecting DIF into four broad categories; 1). *Analyses based on item difficulty*: These approaches compare item difficulty estimates. 2). *Nonparametric approaches*: These procedures use contingency tables, chi-square, and odds ratios. 3). *Item-response-theory-based approaches* which include 1, 2, and 3-parameter IRT analyses. 4). *Other approaches*. These include logistic regression, which also employs a model comparison method, as well as generalizability theory and multifaceted measurement, which are less commonly used in classic DIF studies.

A large range of possible techniques is available; however, only a limited number are currently used. The following section attempts to consider Item Response Theory (IRT)-based models, specifically the Rasch model as an applicable and germane method to present research.

The Rasch Model

IRT is an extension of classical testing theory with mathematical roots which deeply penetrated in psychology and the mathematical basis of IRT has been embedded in the psychological measurement (Ostini & Nering, 2006). Some controversial issues, however, exist in defining the concept of measurement in human science and psychology. Rasch model is mathematically equivalent to the one-parameter logistic (1PL) IRT model, but they developed separately (DeMars, 2010). Controversy surrounds the Rasch model; some specialists believed that the Rasch model and IRT models are structurally different and are used very differently. It is claimed that IRT models are used to "describe and fit data; when fit is poor, the model is adapted or discarded in favor of another model" and, in contrast, 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 involves "model-based measurement in which trait level estimates depend on both the persons' responses and on the properties of the items that were administered" (Embretson & Reise, 2000, p. 13). Furthermore, the test items should not act differently for any specific subgroups of the participants. If an item behaves differently for particular groups, then the validity of the measure for the certain construct decreases; as it is considered as a threat to the test fairness. The Rasch model approach permits investigation of the biased items toward different subgroups and to inspect the construct irrelevant factors (i.e., gender, ethnicity, and academic background) via calculating Differential Item Functioning (DIF) measures.

Besides that, the Rasch model assumptions include unidimensionality and local independence. A unidimensional test consists of items that refer to only one dimension; as DeMars (2010) asserted "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). Wale (2013) mentioned that the assumption of unidimensionality requires "the items function in unison and all non-random variance in the data can be accounted for by person ability and item difficulty" (p. 56). Generally, unidimensionality indicates whether the items makes a single latent trait (θ) (DeMars, 2010).

One aspect regarding unidimensionality deserves caution. Sometimes, responses to 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 both test-taking speed and knowledge (DeMars, 2010).

Another assumption of the Rasch model is local independence. Local independence is "the probability of a test taker responding correctly to a certain item is not dependent on previous responses or the answers given by other individuals to the same item" (Wale, 2013, p. 56). Unidimensionality can be checked via model fit statistics. Besides, unidimensionality and local independence are estimated using fit model statistics; to say a person or an item may be misfitting means the extent to which an intended person and item does not act as the Rasch model would predict.

METHODS

Participants

The participants of the present study (N = 5000) were selected from among the pool of examinees from a population of 20,000 who had taken a recent version of the NUEEFL test in 2015. The participants were randomly selected from the two gender groups (i.e., males and females). The female group included 3335 persons of the total participants and the rest of 1665 examinees were male. The academic background and the age of the participants were not considered in this study.

Instrumentation

The National University Entrance Examination for Foreign Languages (NUEEFL) has a total of 95 items of which 25 are general questions and 70 items come under six subtests: a) Grammar (10 items), b) Vocabulary (15 items), c) Sentence Structure (5 items), d) Language Functions (10 items), e) Cloze Test (15 items), f) Reading Comprehension (15 items).

The NUEEFL test is annually administered to more than 100,000 university applicants to attempt to find the B.A degree in governmental university, specifically in the field of foreign languages. The questions all are in multiple-choice format and are dichotomously scored. The test is time restricted with a dedicated time of 105 minutes. Generally, as a rule in NUEEFL test, guessing is not allowed and the test has included negative score for the wrong responses. It means that a total of three wrong answers will expurgate a correct answer.

The latest version of Winsteps software Version 3.92.1 updated in February 2016 was employed for the data analysis (Linacre, 2016a, b). Winsteps constructs Rasch measures from simple data sets (i.e., usually of persons and items) and applies the dichotomous Rasch model. Pearson-test reliability and item reliability of NUEEFL test were excellent (Pearson r = 0.93 and item r = 1).

Procedure

The NUEEFL is administered annually to a large number of test takers all across Iran. The present study focused on the one aspect of test validity which was assessed through the implementation of the Rasch model. To investigate DIF analysis and apply the Rasch model, the statistical and mathematical assumptions must be met.

Data Analysis

The psychometric properties of the items were estimated using the Winsteps software (Linacre, 2016b). Since the dataset was dichotomous the data were analyzed implementing the Joint Maximum Likelihood Estimation (JMLE) method for estimating the Rasch parameters. In the JMLE formula, the estimate of the Rasch parameter happens when the observed raw score for the parameter matches the expected raw score.

The data-model fit estimated through employing the infit and outfit mean-square values to identify misfit and good-fit items. When it is said a person or an item may be misfitting, it denotes that an intended person and item does not act as the Rasch model would predict (Boone et. al., 2014). The fit estimation checks for the model mis-specifications that can be evaluated in the fit between the model and the data (DeMars, 2010). There are two different fit statistics for persons or items; they are called the weighted (infit) which “weights the square residual by the variance of item”, while the unweighted (outfit) “gives the residual the same weight (i.e., 1)” (Wale, 2013, p. 57). The normal range of acceptable fit for both statistics is between 0.70 and 1.3 (Bond & Fox, 2007; Liu, 2010).

Furthermore, like many IRT models, the Rasch model rests on two basic assumptions: unidimensionality and local independence. The unidimensionality assumption requires that there is only one underlying construct measured by the set of items included in the test. That is, the test measures only one factor. The local independence assumption requires that an examinee’s response to an item does not influence his or her response to any other item. Hence, the items must not give a clue to the correct response for another item.

Unidimensionality was checked through Principal Component Analysis (PCA) in Winsteps. What is required is the fact that there must one dominant factor explaining the shared line of covariance among the items (Hambleton, Swaminathan, & Rogers, 1991). Hence, unidimensionality will hold if the first extracted factor explains a much higher amount of the total variance than that explained by the secondary dimensions. As mentioned before, multiple methods for assessing unidimensionality exist, including the data-model fit statistics. However, studies indicated that these statistics lack the sensitivity required to detect multidimensionality. Hence, it is logical to use Principal Components Analysis (PCA) on the raw data and residuals, in addition to checking the data-model fit.

One approach, first proposed by Smith (2002), for assessing unidimensionality within the Rasch model framework, is the Principal Component Analysis (PCA) based method. This approach aims at assessing whether the items are unidimensional enough as to be treated in practice (Smith, 2002). Principal Component Analysis has the advantage of compressing the data, once patterns have been found in the data. It reduces the number of dimensions, without losing too much of information (Smith, 2002). PCA determines whether the set of items represents a single construct or not. The analysis of dimensionality, based on Smith’s approach, involves a two-step process.

> First, the measurement dimension of the scale was estimated using the Rasch model. The variance associated with this measurement dimension was extracted from the item-response data by computing standardized residuals: (observed - expected)/ (model standard error). Second, a principal component analysis of the standardized residuals was used to determine whether substantial subdimensions existed within the items. If the items measure a single latent dimension as estimated by the Rasch model, then the remaining residual variance should reflect random variation.
>
> McCreary et al., 2013, p. 6

For determining the unidimensionality of questions in the PCA method, conventional criteria were used for judging unidimensionality (Linacre, 2006). Regarding this, Reckase (1979) suggested the following criteria for unidimensionality: a). if the amount of variance explained by measures be > 20%, 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, 2006, p. 272).

According to Smith's approach, at first, the parameters for all questions were estimated, which is called level A in the present research. In the PCA approach, item residuals with loadings of +0.3 or more and −0.3 or less are taken as potential representatives of subdimensions (Hagell, 2014). And in the second round of data analysis, it is attempted to detect the questions with the outfit MNSQ statistic value larger than 1.3. This phase is called level B in which the parameters of the questions were separately estimated. Then, the difference in the difficulty parameter of the questions which were obtained in level B, from the difficulty parameter in level A was estimated. Accordingly, the mean scores of the differences was calculated. Technically, the estimation is called constant correction.

In the next step, the participant's ability parameter were calculated once regarding the entire test (Level A), and once the items calculated separately (Level B). Afterwards, the constant correction value which had been previously calculated was added to the ability parameter (i.e., the Level B). Finally, the difference between the individuals' ability level in the entire test and in the separate items in Level B were calculated through a series of independent t-test for determining the statistical significance and "to compare the two estimates on a person-by-person basis in order to determine the proportion of instances in which the two item sets yield different person measures" (Hagell, 2014, p. 460).

In order to determine the significance of the t-test, the error level of type one ($\alpha = 0.05$) has been modulated. Smith (2002) indicated that if the level of significance of t-test exceeds 5%, the local independency and unidimensionality will be violated.

Moreover, the DIF analysis was examined to test the invariance of measurement. The test developers deploy several quality control or statistical procedures to ensure that the test items are proper and fair for all examinees (Camilli & Penfield, 1997; Holland & Wainer, 2012; Ramsey, 1993).

According to Angoff (1993) an item which shows DIF has different statistical properties in different groups when monitoring for differences in the abilities of groups. DIF is highlighted as an unexpected difference between two groups after matching to measure the underlying ability in the intended item (Camilli, 2006; Wiberg, 2007). Besides, the essential component of DIF analysis in the Rasch model is to compare the item difficulties obtained from the two samples. If the difference in the difficulty estimates is large, then measurement invariance fails and DIF has happened. In the present study the DIF analysis was carried out to test whether the items functioned differently across gender groups.

RESULTS

Data-model Fit Estimation

The Winsteps software normally assessed 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 show a deviation from the unidimensionality, and values less than 1 indicate the overfit in the response patterns with the data-model. The overfit in the model implies the existence of dependency among responses or items.

It is worth mentioning that this statistical index is very sensitive to the sample size. For the sample size with less than 90 people, the model fits with any types of data model, whereas the model does not fit with samples consisting more than 900 people.

Therefore, due to the large sample size in this study and keeping with the valued guidance provided by Linacre (2012), the data-model fit was displayed through MNSQs. The Rasch model offers two indicators of misfit: the infit and outfit mean square indices. Infit is "sensitive to unexpected responses to items near the person's ability level" and outfit discusses "difference between observed and expected responses regardless of how far away the item endorsability is from the person's ability" (McCreary et al., 2013, p. 7). The MNSQs estimates and reports both outfit and infit MNSQs for analyzing the fit of the model.

For both indicators, values between 0.70—1.3 are considered as acceptable or so called good fit values.

Values less than 0.70 indicate outfit, whereas values above 1.3 are a sign of infit. Furthermore, Linacre favors outfit-MNSQs over infit-MNSQs. Hence, in the present research in order to assess data model fit it is decided to consider outfit-MNSQs as a criterion for the outcome interpretations. And, the acceptable values for this index are in the range of 0.70 to 1.3.

In analyzing the model fit estimation, it is required to eliminate the participants with total score of zero. The data were screened for outliers. Besides, the fit indices should be reported for the item calibration. The difficulty estimates for the items, standard errors of item difficulty of estimates, and the infit-MNSQs and the outfit-MNSQs indices are shown in Table 1.

Note that in Table 1 due to space restriction, only the estimation of difficulty parameter and model fit estimations of misfit items are illustrated in descending order (from the most difficult to the least difficult).

Table 1. Item Statistics for Fit Model Estimate and Difficulty Parameter in the Entire Test (Descending Order)

Item	Entry Number	Total Score	Total Count	Measure	Model S.E.	Infit MNSQ	infit ZSTD	Outfit MNSQ	Outfit ZSTD
Q155	80	158	5000	2.45	0.08	1.07	1	1.82	4.8
Q126	51	191	5000	2.23	0.08	1.08	1.2	2.05	6.3
Q137	62	265	5000	1.84	0.07	1.04	0.8	1.32	2.6
Q105	30	285	5000	1.76	0.07	1.19	3.7	3.01	9.9
Q118	43	314	5000	1.64	0.06	1.18	3.6	1.98	7.3
Q166	91	330	5000	1.57	0.06	1.05	1.2	1.49	4.2
Q101	26	335	5000	1.56	0.06	1.26	5.3	3.3	9.9
Q103	28	363	5000	1.46	0.06	1.28	6.1	2.99	9.9
Q158	83	367	5000	1.44	0.06	1.14	3.1	1.46	4.2
Q111	36	392	5000	1.36	0.06	1.12	2.9	1.87	7.4
Q115	40	411	5000	1.3	0.06	1.14	3.5	2.02	8.6
Q133	58	411	5000	1.3	0.06	1.1	2.5	1.77	6.8
Q121	46	459	5000	1.15	0.05	1.38	9.2	2.43	9.9
Q167	92	462	5000	1.14	0.05	0.83	-4.9	0.64	-4.7
Q109	34	463	5000	1.14	0.05	1.24	6	2.03	9.2
Q128	53	496	5000	1.05	0.05	1.2	5.3	1.44	4.7
Q122	47	600	5000	0.79	0.05	1.13	4.2	1.46	5.4
Q156	81	732	5000	0.5	0.04	1.2	6.9	1.41	5.4
Q99	24	821	5000	0.33	0.04	0.83	-7.1	0.69	-5.7
Q84	9	860	5000	0.26	0.04	1.28	9.9	1.6	8.5
Q153	78	874	5000	0.24	0.04	0.78	-9.4	0.59	-8.1
Q149	74	1046	5000	-0.05	0.04	0.75	-9.9	0.57	-9.9
Q108	33	1079	5000	-0.1	0.04	1.17	7.6	1.33	6.1
Q160	85	1151	5000	-0.21	0.04	0.8	-9.9	0.67	-8.1
Q91	16	2076	5000	-1.37	0.03	0.76	-9.9	0.67	-9.9
Q79	4	2513	5000	-1.85	0.03	1.2	9.9	1.36	9.9
Mean		1170.9	5000	0	0.04	1	-0.7	1.14	0.1
P.SD		790.5	0	1.21	0.01	0.13	5.4	0.5	5.3

The first column of Table 1 is the Item Number. The second column shows the Entry Number which provides the order of entering the data in each row. The next column provides Total Score which is total number of correct answers and Total Count is the total number of participants or responses. The fifth column, the measure, reports the difficulty estimates for the items. The range value of difficulty parameter is from

2.45 to -2.99, with mean score of 0, and Standard Deviation (SD) of 1.21. The descending management of item statistics in Table 1 is helpful to arrange the most difficult item Q.155 (Measure = 2.45) and the least difficult item Item Q.87 (Measure = -2.99). Column six provides the Standard Error (SE) of item difficulty estimates.

In the last four columns the infit and outfit statistics are presented. In the following tables the infit-MNSQs values were provided, however; in estimation of data model fit the values of outfit-MNSQs were merely used. As explained before, the acceptable range for both infit and outfit MNSQs is between 0.70 and 1.33. In this study, the outfit- MNSQs indices are all within the acceptable range. However, it appears 26 items were not located in the acceptable range. In Table 1 the range value of the outfit-MNSQs varies from 0.57 to 3.33 which means that some items presented in Table 1 do not fit the model. The investigation of item statistics of outfit-MNSQs reveals that 26 items (27% of items) are not fitted.

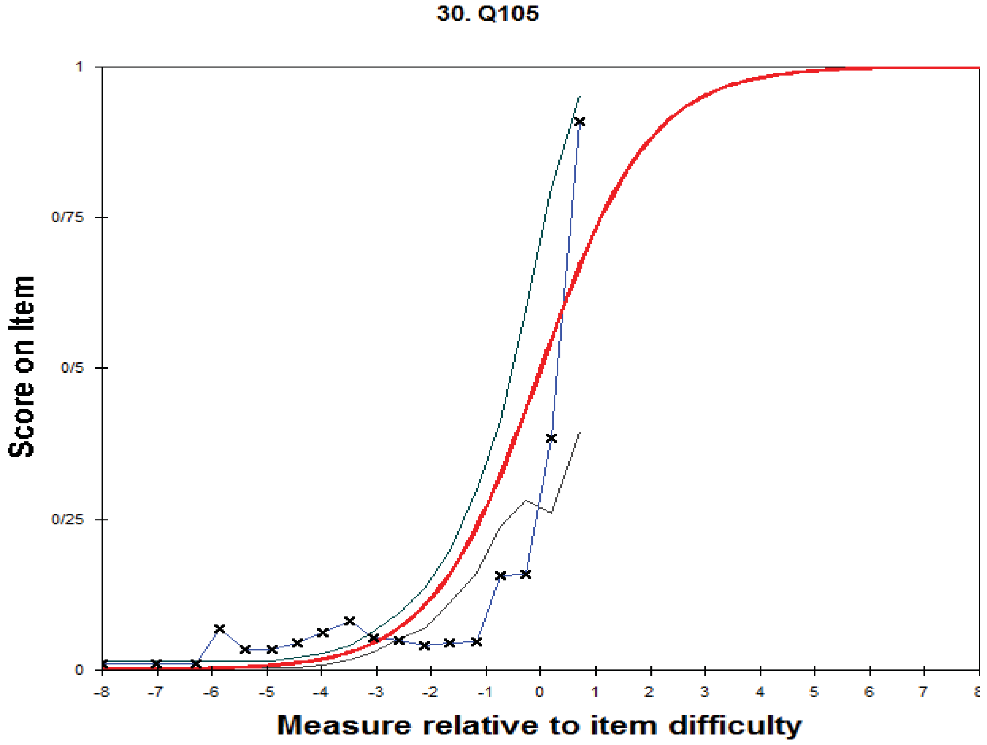

Figure 1. **The ICC Curve for Misfitting Item, Item 105.**

Figure 1 shows the Item Characteristic Curve (ICC) for Item 105. The red curve is the expected ICC. It would be gained if the data fitted the Rasch model. The blue curve is the observed or empirical ICC. The grey line in the outskirt of the red curve is the confidence interval. The confidence intervals are constructed from an estimate and its standard error. The data in present study employed from a large data set, it is evident that the standard error would be very small. The confidence interval become wider if the output of data analysis contains the large standard error.

Figure 1 shows that the empirical ICC for mis-fitting Item 105 has a large deviation from the expected ICC. This fact is reflected in Item 105 with the large outfit-MNSQs value of 3.01. On the contrary, for instance, Figure 2 demonstrates the empirical ICC for good-fitting item (i.e., Item 142). The out-fit MNSQs value (i.e., 0.94) is within the acceptable range.

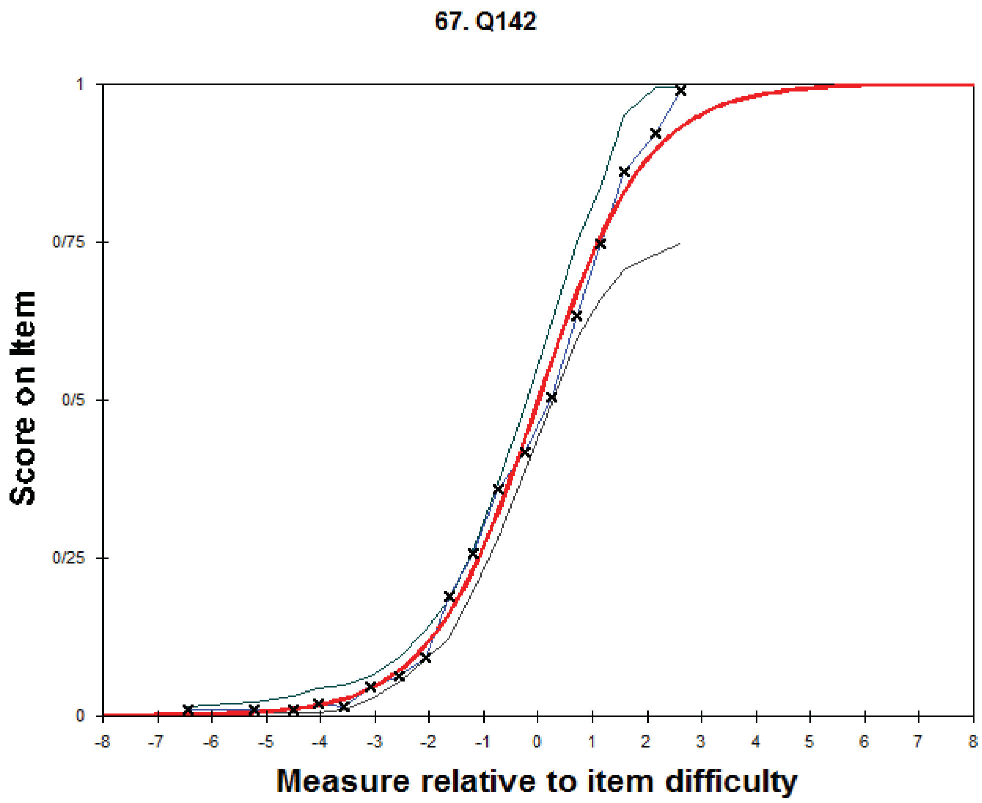

Figure 2. **The ICC Curve for Good-fit Item, Item 142.**

The presence of a large number of misfitting items demonstrates that the data does not fit the model in the NUEEFL. Therefore, the model and its assumptions may be violated. It is possible that the Rasch model unidimensionality assumptions also may not attain the desirable results. Thus, in the next section the results of unidimensionality and local independence will be reported.

Unidimensionality

The unidimensionality assumption requires that there is only one underlying construct measurement by the set of items included in the test. That is, the test measures only one factor. There are multiple methods for assessing unidimensionality, including the data-model fit statistics. However, studies indicated that these statistics do not have the ample sensitivity required to detect multidimensionality. Hence, it is logical to use a Principal Component Analysis (PCA) on the raw data and residuals, in addition to checking the data-model fit. In the current research, the unidimensionality of the test and its items were checked through a Principal Components Analysis (PCA) of residuals and a t-test.

In order to assess dimensionality, PCA of the Rasch residuals was performed. The variance of the measurement dimension is 34.8% with 12.7% of raw variance explained by persons and 22.1% raw variance explained by items. The results showed that it is larger than the requirement of 20%, demonstrating a unidimentional trait of the data (Reckase, 1979).

The first, second, third, fourth, and fifth unexplained variance accounted for eigenvalues are 3.4, 2.5, 2.2, 1.9, and 1.7 which were good by referring the criteria. The results of the data analysis suggested that the unidimensionality is hold across the whole test (See Table 2 & 3).

Table 2. PCA Analysis

Variance in Eigenvalue units	Eigenvalue	Observed	Expected
Raw variance explained by measures	50.7068	34.8%	33.7%
Raw variance explained by person	18.5454	12.7%	12.3%
Raw variance explained by items	32.1614	22.1%	21.4%
Raw unexplained variance	95.0000	65.2%	66.3%
Total raw variance in observations	145.7068	100.0%	100.0%

Table 3. Contrast in Eigenvalue Units

Contrast in Eigenvalue units	Eigenvalue	Observed	Expected
Unexplained variance in 1st contrast	3.4662	2.4%	3.6%
Unexplained variance in 2nd contrast	2.5397	1.7%	2.7%
Unexplained variance in 3rd contrast	2.2560	1.5%	2.4%
Unexplained variance in 4th contrast	1.9438	1.3%	2.0%
Unexplained variance in 5th contrast	1.7076	1.2%	1.8%
Raw unexplained variance (total)	95.0000	100.0%	100.0%

Local independence was examined through checking all the abilities in order to identify whether the responses to items could be independent of each other (Pae, 2011). As for local independence, in assessing the t-test statistics the outfit-MNSQs was examined. The results showed that 20 items for this statistic exceeded 1.3. These 20 items are considered as those which confirm the unidimensionality assumption in data analysis.

The steps of t-test calculation were conducted. Note that the measurement of local independence was based on Smith's approach (2002). The total sum of difference between difficulty parameter of these 20 items gained from level A and B is equal with -1.12. The constant correction value is -0.056 which is obtained through dividing -1.12 by the number of items (20 items). By adding the constant correction value to the ability parameter of participants within questions of level B. Then, it is attempted to examine the significance level of t-test and to modify the significance level.

The Student's t-statistics on 20 items revealed that there were no significant different among t-test results. Thus, it is concluded that the local independence assumption is strongly accepted in the entire NUEEFL test. To sum, regarding the results both Rasch assumptions (i.e., unidimensionality and local independence) are hold in the whole test.

Differential Item Functioning

The next step in data analysis was the DIF analysis. The ability and difficulty estimates in the Rasch model are assumed to be invariant. The statistical procedure aims at identifying items with different statistical features across certain groups of examinees.

In this study, DIF analysis was investigated for the gender groups and for the NUEFFL items. For DIF Analysis in a Rasch context, both magnitude of the difference in logit units between the groups and statistical significance of the difference should be considered (Linacre, 2016a). The magnitude of the DIF value should be at least 0.5 logits, indicating the comparison between differences in difficulty of the items for one group to the difficulty level of the same items for the other group (Linacre, 2016a).

In this stage of this study, DIF analysis was tested between two groups of males and females. In order to examine the invariance, the difference between the DIF analysis of two groups of males and females through testing the t-test of the statistical significance of the data was investigated. For statistically significant DIF, the probability of such difference (0.5 logits or larger), occurring as a random accident, should be ≤ 0.05. This indicates that the probability of such difference happens when there is no systematic item bias in the test items (Linacre, 2016a).

Beside, considering that statistical significance tests are affected by sample size and due to the large size of the study groups in present research, it needs at least 0.5 logits for DIF to be noticeable. For instance, if the difficulty of an item in both groups of males and females has 0.5 logits difference, that specific item will be considered as a DIF item. Note that, due to space limitation, only DIF-flagged items appear in Table 4.

Table 4. DIF-flagged items in the NUEEFL test

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
Q76	M	-0.87	F	-0.67	-0.2	-2.75	INF	0.0059
Q77	M	-1.95	F	-2.11	0.15	2.19	INF	0.0289
Q80	M	-2.5	F	-2.08	-0.42	-5.85	INF	0.0000
Q85	M	-0.16	F	0	-0.16	-1.96	INF	0.0498
Q86	M	-1.17	F	-1.02	-0.15	-2.05	INF	0.0400
Q90	M	0.29	F	0.81	-0.52	-5.6	INF	0.0000
Q91	M	-1.48	F	-1.31	-0.17	-2.47	INF	0.0137
Q94	M	-0.99	F	-0.81	-0.18	-2.5	INF	0.0123
Q95	M	-1.08	F	-0.72	-0.35	-4.83	INF	0.0000
Q97	M	-0.41	F	-0.12	-0.29	-3.59	INF	0.0003
Q99	M	0.14	F	0.45	-0.31	-3.52	INF	0.0004
Q103	M	1.62	F	1.36	0.26	2.1	INF	0.0362
Q104	M	0.55	F	0.35	0.21	2.22	INF	0.0263
Q105	M	2.01	F	1.61	0.4	2.89	INF	0.0039
Q106	M	-0.46	F	-0.09	-0.37	-4.66	INF	0.0000
Q107	M	-0.4	F	-0.63	0.23	2.93	INF	0.0034
Q108	M	0.03	F	-0.17	0.2	2.42	INF	0.0157
Q111	M	1.62	F	1.21	0.41	3.35	INF	0.0008
Q113	M	-0.48	F	-0.07	-0.41	-5.15	INF	0.0000
Q117	M	-1.19	F	-0.98	-0.21	-2.93	INF	0.0034
Q119	M	1.31	F	1.67	-0.36	-2.99	INF	0.0028
Q121	M	1.52	F	0.96	0.55	4.77	INF	0.0000
Q122	M	0.96	F	0.7	0.26	2.58	INF	0.0100
Q125	M	0.36	F	0.81	-0.44	-4.71	INF	0.0000

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
Q128	M	1.2	F	0.96	0.24	2.2	INF	0.0279
Q130	M	0.39	F	0.18	0.21	2.37	INF	0.0179
Q131	M	-0.18	F	-0.41	0.22	2.78	INF	0.0054
Q132	M	-1.38	F	-1.7	0.32	4.55	INF	0.0000
Q135	M	-0.96	F	-1.42	0.46	6.43	INF	0.0000
Q138	M	-0.86	F	-1.07	0.21	2.84	INF	0.0046
Q141	M	-1.81	F	-2.02	0.21	2.95	INF	0.0032
Q142	M	0.08	F	-0.12	0.2	2.35	INF	0.0187
Q143	M	-0.49	F	-0.3	-0.19	-2.46	INF	0.0140
Q145	M	-0.9	F	-1.14	0.24	3.3	INF	0.0010
Q147	M	1.74	F	1.33	0.41	3.28	INF	0.0011
Q157	M	-0.04	F	0.24	-0.28	-3.3	INF	0.0010
Q159	M	0.47	F	0.97	-0.5	-5.12	INF	0.0000
Q163	M	0.05	F	-0.18	0.22	2.69	INF	0.0072
Q167	M	0.96	F	1.27	-0.3	-2.8	INF	0.0052
Q168	M	-0.42	F	-0.89	0.47	6.22	INF	0.0000

Note. M = Male; F = Female; INF = Infinity

The DIF analysis between groups of male and female was investigated. The results of this analysis are shown in Table 4. The results show that among 95 items, 40 items exhibit DIF.

The difficulty level of items between males and females was variant. Hence, it is concluded that the invariability of questions in gender group is not accepted. The null hypothesis which stated the participants' gender is not a source of DIF in NUEEFL is rejected. Given significance DIF within the Rasch model in gender group, the NUEEFL appeared not to be a DIF-free person estimates test. Hence, it is concluded that there is significant difference between males and female in answering the NUEEFL test. And, the NUEEFL test is not fair to all male and female participants.

DISCUSSION AND CONCLUSIONS

The present study aimed at investigating the interaction of person abilities and item difficulties of the high-stakes test of university entrance exam in Iran (i.e., NUEEFL). In particular, results from the Rasch model and DIF analysis were compared to see whether evidence of differential functioning would be found in data analysis.

The descriptive statistics revealed that there was significant difference in the overall test. Hence, the hypothesis that the data fit the Rasch model was not supported. The results of data-model fit revealed that 26 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 total 95 items were not located in acceptable range value of 0.70 to 1.30. There are many misfitting items and items were not fitted with the model.

A concern about the dimensionality of the NUEEFL test suggests that a calibration of test of Rasch model in Winsteps software revealed that there are many misfitting items. The dimensionality was detected through the Principal Component Analysis (PCA) on the raw data and residuals. The amount of variance explained by the different components in the data was 34.8% (eigenvalue 50.70) which is larger than 20% as to be indicative of unidimensionality. The results of the data analysis suggested that the unidimensionality is hold across the whole test.

In the case of local independence, a series of t-test was performed. The result of outfit-MNSQs showed that 20 items are larger than the intended criteria (i.e., 1.3). It is concluded that the local independence assumption is strongly accepted in the entire NUEEFL test.

DIF analysis confirmed a different probability of endorsing the test items across the gender groups. Based on the DIF results, it is interpretable that out of the 95 items, 40 items displayed 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.

Additionally, there is an ongoing interest in comparing cultural, ethnic, or gender groups. DIF studies are essential in testing programs with high stakes. Furthermore, possible gender and/or ethnicity bias could negatively impact one or more groups in a construct-irrelevant manner. In fact, the test administrators attempt in making a perfectly fair testing battery; however the dearth of research on NUEEFL test raised questions regarding the fairness of this national test.

The NUEEFL, which is given to thousands of individuals annually, is used as a gate-keeping test for entering higher education in Iran. In line with the main purpose of this research, DIF analysis across gender groups was investigated. DIF analysis rejected a similar probability of endorsing the test items across the gender groups. The results of the study indicated that 40 items out of the 95 items of NUEEFL test exhibited DIF. This suggests that NUEEFL test scores are not free of construct-irrelevant variance.

It is worth mentioning that the results of the present study are not consistent with Karami's (2015) study from the aspect of dimensionality of NUEEFL, in which the multidimensionality in the whole test and among sub-tests had been proven.

According to Camilli (2006) DIF analysis mainly focuses on the performance of two or more different groups. Therefore, such analysis is unable to disclose the existence of bias against different individuals. Moreover, this study directs to this point that due to the assessment of the high-stakes tests such as NUEEFL is required to consider the performance of individuals besides the groups' performance.

Also, in NUEEFL the performance of individuals should be examined. It is proposed that the students should be clustered based on their weak points detected through the University Entrance Exam. Then, after acceptance by universities, the student vulnerability in any of the subtests should be reported to the targeted university. Specifically, this valuable report can be of help to the students because the universities can then offer them some prerequisite English courses.

It should be noted that in the present research only a receptive skill of reading which mainly focused on reading comprehension, vocabulary, and grammar, was examined. The study excluded listening, writing, and speaking skills. In listening comprehension, for instance, females were found to have an advantage compared to males (Boyle, 1987; Cole, 1997). The results of DIF analysis in the present study indicated that there is DIF between males and females and the findings from this study are different from with the results of Ryan and Bachman (1992) who found no gender difference in any of TOEFL subtests.

The major contribution of this study is to the field of language testing with providing empirical evidence for the interpretation of NUEEFL items and scores via a comprehensive investigation of the item fit, dimensionality, and the detection of biased items. Evaluating the constructs and test fairness for the university entrance exam, specifically the fairness study in the NUEEFL is essential because it is currently used for gauging high school knowledge by the National Organization for Educational Testing (NOET).

REFERENCES

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.

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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch Analysis in the Human Sciences. Dordrecht: Springer Science, and Business Media.

Boyle, J. (1987). Sex differences in listening vocabulary. *Language Learning, 37*(2), 273-284.

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.

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

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

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

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.

Hagell, P. (2014). Testing rating scale unidimensionality using the Principal Component Analysis (PCA)/*t*-Test Protocol with the Rasch Model: The primacy of theory over statistics. *Open Journal of Statistics, 4,* 456-465. doi: 10.4236/ojs.2014.46044.

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

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

Karami, H. (2013). The quest for fairness in language testing. *Educational Research and Evaluation*, *19*(2–3), 158–169.

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

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

Linacre, J. M. (2010). Rasch measurement: Transactions of Rasch measurement *SIG American Educational Research Association, 24*(3)*,* pp. 1289–1300.

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

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

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

Liu, X. (2010). Using and developing measurement instruments in Science education: A Rasch Modeling Approach. Charlotte, NC: Information Age (pp. 224- 228).

McCreary, L. L., Conrad, K. M., Conrad, K. J., Scott, C. K., Funk, R. R., & Dennis, M. L. (2013). Using the Rasch measurement model in psychometric analysis of the family effectiveness measure. *Nursing Research, 62*(3), 149-159.

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

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

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

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.

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.

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

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.

Sireci, S. G., & Rios, J. A., (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation,* 9(2–30), 170–187, doi 10.1080/13803611.2013.767621.

Smith Jr., E. V. (2002) Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals. *Journal of Applied Measurement*, 3, 205-231.

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.

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

Zwick, R., Ye, L., & Isham, S. (2013). *An investigation of the efficacy of Criterion Refinement Procedures in Mantel-Haenszel DIF Analysis* (pp.1-29). Princeton, NJ: Educational Testing Service.