ATU PRESS

ISSUES IN LANGUAGE TEACHING (ILT)
VOL. 10, NO. 1, 147-170, June 2021
https://doi.org/10.22054/ilt.2021.60283.588 Document Type: Research paper

Examining the Fairness of the University Entrance Exam: A Latent Class Analysis Approach to Differential Item Functioning

Sayyed Mohammad Alavi
Professor of Applied Linguistics, University of Tehran, Tehran, Iran

Hossein Karami
Assistant Professor of Applied Linguistics, University of Tehran, Tehran, Iran

Mohammad Hossein Kouhpaeenejad*
Ph.D. Candidate of TEFL, University of Tehran, Tehran, Iran

Received: May 13, 2021**; Accepted:** June 29, 2021

Abstract

Measurement has been ubiquitous in all areas of education for at least a century. Various methods have been suggested to examine the fairness of education tests, especially in high-stakes contexts. The present study has adopted the newly proposed ecological approach to differential item functioning (DIF) to investigate the fairness of the Iranian nationwide university entrance exam. To this end, the actual data from an administration of the test were obtained and analyzed through both traditional logistic regression and latent class analysis (LCA) techniques. The initial DIF analysis through logistic regression revealed that 19 items (out of 70) showed either a uniform or non-uniform DIF. Further examination of the sample through LCA showed that the sample is not homogeneous. LCA class enumeration revealed that three classes can be identified in the sample. DIF analysis for separate latent classes showed that three serious differences in the number of DIF items identified in each latent class ranging from zero items in latent class 3 to 43 items in latent class 2. The inclusion of the covariates in the model also showed that latent class membership could be significantly predicted from high school GPA, the field of study, and the acceptance quota. It is argued that the fairness of the test might be under question. The implications of the findings for the validity of the test are discussed in detail.

***Keywords*:** Fairness; Differential item functioning; Bias concept of validity

***Corresponding author's email:** mh.kouhpaee@ut.ac.ir

INTRODUCTION

The concern with test fairness dates back to the beginning of the twentieth century. Although it might be difficult to define fairness, it may be considered as the opposite of bias. That is, a test which is not biased is said to be fair. Various statistical methods have been introduced to address test and item bias. One such technique is differential item functioning (DIF). In fact, bias and DIF were once used interchangeably to denote the same concept. However, "because of its semantic association with societal issues like discrimination, another term was coined for more technical analyses of test items: differential item functioning" McNamara and Roever (2006, p. 82).

However, when it comes to appraising exactly what DIF denotes in practice and the relationship between DIF and bias, some researchers seem to ignore this key point about DIF. For instance, McNamara and Roever (2006) discuss the implications of DIF for test fairness and bias. They rightly point out that DIF is not necessarily an indication of bias. They go on to give examples of the cases in which DIF is not a sign of test bias. One of their examples is related to cases where "items function differently for two groups of test-takers simply because the two groups differ in their ability" (p. 84-85). It is not clear how this would show up as DIF if conditioning on an estimate of ability is already done.

The vast majority of the DIF studies reported in the literature are based on an observed grouping variable such as gender or academic background. The problem with such studies is that they address only one possible source of DIF. The lack of DIF in such studies does not mean that there is no DIF: only DIF arising from that particular grouping variable does not exist. The point is we cannot predict apriori which of these independent variables can cause DIF before actually examining all of them. More recent approaches to DIF analysis, however, are based on a latent categorization of the respondents. This study has examined DIF in the university entrance examination based on the framework proposed by Zumbo et al. (2015).

Before discussing the details of the study, however, a review of the relevant literature is provided in the next section.

LITERATURE REVIEW

The advent of solid bias studies in psychometrics seems to be in the 1960s. Osterlind and Everson (2009) argue that the early focus on bias studies received an impetus with the enactment of the Civil Rights Act in 1964. Cole also highlights the importance of the new law for bias studies, stating that" "test and item bias concerns in their modern form grew out of this era, were responses to it, were influenced by it, and took their role as a standard part of the enterprise because of it" (Cole, 1993, p. 25).

The Civil Rights Act in the United States brought with it new challenges. Test developers and users were required to produce tests that cherished equality and equity among test takers of various backgrounds. This emphasis on test equity was not out of mere interest in psychometrics and educational measurement. Rather, the social demand for equity left no other option for measurement professionals.

It should also be noted that McNamara and Roever (2006) trace the history of bias studies to the early years of the 20th century. This line of research was mostly concerned with group comparisons on mostly IQ tests and exploited various statistical techniques to investigate bias. Interestingly, the majority of these studies made the groupings based on the socioeconomic status of the examinees. Specifically, they compared examinees coming from different socioeconomic backgrounds. This is unlike the bias studies done in the second half of the twentieth century where there was a shift of attention from social variables to test-taker background variables such as gender and race. This shift of focus is surprising when we consider the social demands for equity in education and educational assessment during the post-Civil Rights Act era.

Bias and DIF studies gained momentum during the 1970s from the Golden Rule Settlement. The Golden Rule Insurance Company sued the

Illinois Department of Insurance and Educational Testing Service (ETS) for "alleged racial bias in the licensing test for insurance agents" (McNamara & Roever, 2006). The case was initially dismissed by the primary court. However, the insurance company refiled and the circuit court ruled in favor of the Golden Rule Company.

Such a verdict had serious repercussions for the testing industry. It highlights the role of negative social consequences of test scores (Messick, 1989). In addition, testing corporations faced the risks of imprudent developments of tests and the kind of legal consequences they could bring about.

DIF analyses have been increasingly applied since the 1970s. Various methods of DIF detection techniques have also been introduced in the literature. Zumbo (2007) divides the history of DIF analysis into three generations. This is discussed in the next section.

Zumbo (2007) has argued that there have been three generations of DIF analysis. As he points out, these generations are not meant to indicate "distinct historical periods" there are certain commonalities among the three generations.

During the first generation, the term bias was more frequently used to indicate what is now called DIF. The concern with fairness and equity was a serious concern in high-states testing contexts. Any differential performance on such tests would bring about a discussion about the possible bias in the test. Most of the time, the group which was suspected to be disfavored by the test would be called the focal group and the group of examinees against whom the focal group was compared was dubbed the reference group.

Zumbo (2007) argues that the move to the second generation of DIF analysis was marked by two notable changes. First, the more neutral term differential item functioning was introduced to be used instead of the more commonly used term bias. This was significant in that bias has certain connotations and cannot be neutrally used for denoting differential performance of groups that is not construct-irrelevant.

The second significant change in this generation was the separation of

bias from impact. During this generation, bias was used to refer to the differential performance of different groups of examinees which was not due to true differences in ability level. Hence, a factor other than the construct measured by the test was causing the difference. Impact, on the other hand, referred to cases where two groups of examinees received different scores and their different performance on the item was due to true differences in underlying ability measured by the item.

Since the third generation of DIF analysis has its roots in the second generation, an additional key feature of the second generation must be noted here. As Zumbo (2007) also explains, there have three broad categories of methods for identifying DIF items. The first group of models is based on contingency tables and regression analysis. Considering the definition of DIF as one group of examinees giving a correct response to an item more often compared to equally knowledgeable members of other groups of examinees, it follows that the examinees must be matched for their ability levels before DIF analysis is done. This matching or conditioning is the basis of the first group of DIF detecting techniques. Hence, the common thread running through the models in the first category is the fact that conditioning on ability level should omit all the differences in the frequency of giving a correct response in the two given groups. Any remaining differences would count as DIF. Logistic regression which has been one of the most frequently used methods of DIF detection belongs to this category. This method will be explained in adequate detail below as it is at the heart of DIF analysis in the present study.

The second group of models is based on item response theory (IRT). There is no matching or conditioning in this category as these models assume that the two groups of examinees have the same ability distribution (Zumbo, 2007). Specifically, these models estimate the differences in the item characteristic curves (ICCs) for the two groups. If there is no difference in the ICCs for the two groups, then it is assumed that there is no DIF. On the other hand, if the ICCs are different, the item is flagged for DIF. The ICCs can be different in two ways. If only the difficulty parameter is

different, it is stated that the item has uniform DIF. The item is said to show non-uniform DIF if the ICCs differ in item discrimination (in which case the ICCs would cross).

The last group of DIF detection models is multidimensional models. The assumption behind the multidimensional models is that most of the time, the items do not measure a single construct. That is, in addition to the primary dimension which is measured by the item, there might also be some secondary dimensions that lead to variance in the score above the already existing variance which is due to the primary construct. Hence, these secondary dimensions might be responsible for DIF. this view of DIF is also evident from some definitions of bias. For example, McNamara and Roever (2006, p. 82) argue that:

Another way to look at this is to consider bias a factor that makes a unidimensional test multidimensional: The test measures something in addition to what it is intended to measure, and the result is a confound of two measurements.

Shealy and Stout's (1993) simultaneous item bias tests or SIBTEST DIF detection methods are possibly the most well-known models in this category. The first step in the SIBTEST model is to identify two subsets of items: valid item subsets and suspected subsets. The suspected items are selected a priori based on the existing literature and extant theories. The interesting point about these models is the fact that they take a confirmatory approach to DIF detecting. Such an approach decreases the burden of explaining why DIF has occurred which is an arduous task in DIF analysis (Alavi & Karami, 2010).

The search for genuine causes of DIF marks a significant point in the history of DIF analysis. The multidimensional models are actually the beginning of offering formal models for identifying causes of DIF. However, Zumbo (2007) notes that the focus of these models is still on factors that pertain to features of the items themselves and that the multidimensional approach "places the source of DIF in the test structure" (p. 229). It must be noted here that the concern with identifying the sources

of DIF has been around from possibly the beginning of DIF analysis. This has been important because if the true causes of DIF are not known, at the end of the day we do not know if we are facing bias or impact.

Zumbo (2007) gives the following account of the third generation of DIF analysis:

[T]he third generation of DIF is most clearly characterized as conceiving of DIF as occurring because of some characteristic of the test item and/or testing situation that is not relevant to the underlying ability of interest (and hence the test purpose). By adding testing situation" to the possible reasons for DIF that have dominated the first two generations of DIF (including the multidimensional model), one greatly expands DIF praxis and theorizing to matters beyond the test structure (and hence multidimensionality) itself, hence moving beyond the multidimensional model of DIF. (p. 229)

Hence, the potential sets of factors that can bring about DIF are not limited to the features of the test and the items at hand. Any other factor in the social context and any variables related to the examinee's background can be taken into account.

Although a plethora of research studies have been conducted to examine various aspects of DIF, there seems to be a major gap in the literature. There have been two broad types of DIF studies. The first type of DIF study has addressed aspects of the theoretical and methodological issues in general. On the other hand, the second group of DIF studies has been more practical in nature and has addressed DIF in specific testing contests and specific measurement tools.

One of the major problems in DIF analysis has been devising a mechanism for identifying the source of DIF when an item is flagged. In the absence of such a mechanism, any explanation would be subjective. In fact, there is evidence that explanations of the sources of DIF can be totally idiosyncratic. For instance, Bond (1993) recites a personal experience of running a DIF analysis and trying with one of his students to identify sources of DIF. After justifying the source of DIF for every item, they found out that they had made a mistake falsely taking non-DIF items as DIF items.

The next step was to justify the sources of DIF for the new set of items. The speculations seemed equally plausible. The bottom line is the speculations about sources of DIF are only speculations. They cannot be taken as proven facts.

In a similar vein, Alavi and Karami (2010) provided empirical evidence on the "ad hoc" nature of the explanations offered for DIF. They conducted a study with a group of language testing experts and asked them to offer explanations for why they thought DIF had occurred in six items. Out of the six items judged by the experts, two items had in fact shown DIF in favor of the given group, two items had disfavored that group and two items had not shown DIF at all. However, the experts were told that all these items favored the group. The results revealed that the experts resorted to all sorts of theorizing to explain the DIF even if some of them were farfetched. Taking into account the results of their study, Alavi and Karami (2010, p. 14) concluded the interpretations of sources of DIF are mostly ad hoc.

Hence, it is clear that there is a dire need in DIF analysis to devise a mechanism to render the explanations of DIF analysis more plausible and objective. This is essential because, as Alavi and Karami (2010) argue, there is little value to the explanations offered for sources of DIF if those explanations are not based on empirical evidence.

The second problem in DIF analysis is related to the way the classification of the examinees into different groups is done. This classification is most frequently done on the basis of an observed background variable such as gender, ethnicity, academic background, and so on. The problem is that any DIF analysis is usually focused on just one of these variables. However, these variables do not necessarily result in different response rates in all contexts. For example, gender might prove to be very influential in one testing context and may not have any effect in another context. The bottom line is that in each testing situation it is known a priori which of these variables is causing the responses to be different.

One solution to this problem is to apply latent class analysis (LCA) (Wang & Wang, 2012). LCA is an approach to classifying the examinees

into latent classes or groups. The classification of the test takers is based on inspection of the observed response patterns of the examinees.

Specifically, the examinees with similar response patterns are categorized into the same groups. LCA can be beneficially applied to DIF analysis. Unlike observed-variable based DIF techniques, the LCA approach to DIF analysis has the advantage of first identifying the different latent groupings that genuinely exist and then see if the equally knowledgeable examinees from the latent classes have different probabilities of correctly responding to an item (i.e., that DIF exists for these groups).

PURPOSE OF THE STUDY

Zumbo et al. (2015) have recently proposed a new method for DIF analysis. Their model is in fact a blend of the traditional logistic regression (LR) model of DIF analysis and latent class analysis (LCA). The latent class model aims to overcome the problems with DIF analyses delineated earlier. By combining the LCA model and the conventional logistic regression model, they provide a solution to the problem of randomly selecting an observed variable for DIF analysis which may not be creating a differential performance in a given testing situation. Second, they formally add covariates or explanatory variables to their model so that the source of DIF may be more objectively identified.

The present study aimed to fill the existing gap in the DIF literature. Specifically, the study applied the model proposed by Zumbo et al. (2015) to detect DIF items in the National University Entrance Exam for Foreign Languages (NUEEFL). The main grouping variable for DIF analysis was the gender of examinees. Other examinee background variables were added to the model to see if they can predict the latent classes.

It should be noted here that a number of DIF studies have been conducted to examine the existence of bias in the NUEEFL test (e.g., Barati & Ahmadi, 2010; Mehrazmay, 2012). The problem with these studies again stems from the fact that they ignore the overall ecological context of the test.

Taking into account the purpose of the study, the following research questions have been addressed:

1. Does the LR DIF analysis identify any gender-related DIF items in the NUEEFL test?
2. How many latent classes exist among the test takers who have taken the NUEEFL test?
3. Is there any noticeable overlap between the latent classes and the background variables of the examinees?
4. Do the latent classes predict the existence of the uniform and non-uniform DIF?
5. Does any of the examinee's background variables (i.e., the covariates) significantly predict the latent classes?

METHOD

Participants

The participants of the present study will be in fact a subset of the test takers who have taken the NUEEFL in 2014. From among the total population of examinees, a group of 5000 were randomly selected to run the DIF analysis. The researcher did his best to obtain the complete data set from the NOET such that every information (except for identifying information) about the examinees can be obtained to render the results of the DIF analysis more dependable. Due to the nature of the current study, the more the number of the examinees' variables entered into the mode, the more dependable the results of the DIF analysis would be.

In this study, four variables were selected as the independent variables: Gender, Field of Study, Quota, and Grade Point Average (GPA). The National Organization for Educational Testing (NOET) which is responsible for the development and administration of the test provided us with a sample of 5000 participants. Among the participants, 66.8 percent (N=3338) are female and 33.2 percent (N=1662) are male. However, after considering other variables such as Quota, 4769 participants remained.

Instrumentation

The data for the current study comes from an administration of the NUEEFL in 2014. The test comprises 70 multiple-choice items which are all dichotomous scores. The items come in six subtests in the following order:

Grammar (10 items)
Vocabulary (15 items)
Sentence structure (5 items)
Language functions (10 items)
Cloze test (10 items)
Reading comprehension (20 items)

A correction for guessing is applied in order to discourage the examinees from resorting to random guessing. The correction for guessing is done through the following formula:

$$S = R - \frac{W}{K - 1}$$

where S is the corrected score, R is the number of correct responses, W is the number of incorrect responses, and k is the number of options. The time limit for taking the entire set of items is 105 minutes. The examinees can go back and forth through the items and there are no time limits on separate sections of the test. Normally, a total score is reported for the whole test and the test takers' performance on each subset of the items is not announced by the NOET. Hence, low performance on a given subtest can be compensated for by better performance on another subtest. The NOET allows no missing data and all missing responses are counted as incorrect.

Data Analysis

The main software utilized for doing the data analysis will be the Mplus software (Muthén & Muthén 1998–2012). The data analysis will be based on the model proposed by Zumbo et al. (2015). There are four steps in their model.

In the first step, the traditional logistic regression model will be used to identify DIF items working differently for males and females. The results of this step can later be compared with the results of the final step to see whether there is any difference between logistic regression and LCA in identifying DIF items.

In the second step, class enumeration is done through latent class analysis. Since the number of classes is not a parameter in the model, successive LCAs must be done each time adding one more latent class (de Ayala, & Santiago, 2017; Rindskopf, 2009). Various fit indices are then used to compare the relative fits of the model with a different number of latent classes. The model with the best fit is then selected and the number of latent classes is determined.

In the third step, covariates were added to the model to predict the latent classes. It must be noted that although there are no theoretical limits on the number of covariates that can be added to the model, the data which was available by the NOET included only a few background variables. Not all of these variables could be used in the study. For instance, although it was specified which provinces the students belonged to, only a few students came from some of these provinces. The scant data could undermine the veracity of the data analysis results. As explained in earlier sections, there were four independent variables in this study: Gender, Field of Study, Quota, and Grade Point Average (GPA).

Multiple covariates are examined at the same time in the fourth step. The interpretation of the covariates in this stage would be different from Step three. The difference between these two steps is like the difference between simple and multiple regression analysis. When there is more than one independent variable (as in multiple regression), the regression coefficient shows the contribution of this independent variable when the effect of all other independent variables is controlled for. Hence, we can obtain a better image of the unique contribution of each independent variable. Step 4 in this study had the same objective and aimed at examining the unique contribution of each of the independent variables when all other

independent variables were held constant.

It should be noted that since there were only four covariates in the present study (compared with 19 covariates in Zumbo et al., 2015), steps 3 and 4 were merged. This certainly does not have any effect on the outcome of the study.

A critical part of latent class analysis, on which the current approach to DIF analysis is based, is determining the number of latent classes. Unlike the traditional approaches such as analysis of variance where the number of groups is predetermined, the number of latent classes in LCA is not known prior to the analysis. LCA assumes heterogeneity in the population of interest (unlike regression analysis for instance) and captures this heterogeneity by discovering the unobserved groupings among the population.

Several fit indices were utilized to compare the relative fit of successive models. Wang and Wang (2012) suggest the following fit indices:

1. Information criterion indices, such as Akaike Information Criterion (AIC), consistent AIC (CAIC), sample-size adjusted CAIC (ACAIC), Bayesian Information Criterion (BIC), and adjusted BIC (ABIC).
2. Lo–Mendell–Rubin likelihood ratio (LMR LR) test (Lo, Mendell, & Rubin, 2001), and adjusted LMR LR (ALMR LR) test.
3. Bootstrap likelihood ratio test (BLRT).

As for the information criterion indices, lower values of these indices indicate better fit. In this study, models with smaller values were selected as models with a better fit. Wang and Wang (2012) argue that a significant value of the LMR LR index indicates that the addition of one more latent class to the model results in significant improvement model fit. A non-significant value, on the other hand, shows that the addition of the new latent class does not improve model fit and, hence, is not justified.

The Mplus software reports a P-value for the BLRT test. It is interpreted in the same way as that of the LMR LR test. Wang and Wang (2012) argue that from among all these indices, BIC and BLRT have the

best performance in determining model fit. In this study, model fit was based on an inspection of as many fit indices as possible to render the findings of the data analysis more dependable.

RESULTS

Logistic Regression

The first step in data analysis was to run a binary logistic regression to identify differentially functioning items across males and females. As Zumbo et al. (2015) argue, the purpose of this step is to provide a criterion against which the results of the latent class DIF analysis can be compared. The formula for DIF analysis through logistic regression is presented below:

$$\text{Ln}\left(\frac{P_{mi}}{1-P_{mi}}\right) = b_0 + b_1 total + b_2 gender + b_3(total \times gender)$$

where b_0 is the intercept, $b_1 total$ is the effect of the total score on the test which acts as the conditioning variable in this study, $b_2 gender$ shows the effect of gender, and finally $b_3(total \times gender)$ is the ability by gender interaction effect. If total score alone can predict item performance, then there is no DIF. If gender also adds to the predictive power of the total score, then we have uniform DIF (where members of one group uniformly outperform equally competent members of another group). Finally, if the interaction between total score and gender is also significant, then we have non-uniform DIF (where members of one group outperform equally competent members of another group up to an ability level and then the direction is reversed).

A close inspection the results revealed that 19 items were flagged for DIF. It appeared that 7 items (i.e., 6, 12, 16, 18, 20, 42, and 50) showed significant uniform DIF. In addition, 14 items (i.e., 2, 4, 7, 13, 14, 35, 36, 38, 39, 41, 42, 45, 50, and 61) were flagged for significant non-uniform DIF.

Latent Class Analysis Enumeration

The second phase of the study was latent class analysis (LCA). The purpose of the LCA was to identify the latent groups (or classes) of participants. This is technically called latent class enumeration in the LCA literature (see Wang & Wang, 2012). The difficulty in determining the number of classes to extract lies in the criteria used for judging model adequacy. In other modeling frameworks such as structural equation modeling (SEM), there are well-researched and frequently used criteria for model selection. In LCA, chi-square distribution is not approximated properly due to the sparse nature of the data (Geiser, 2013). Hence, chi-square-based indices cannot be readily applied for model evaluation. Normally, various models with an increasing number of latent classes are hypothesized. The analysis is stopped when adding another latent class does not result in significant improvement in model fit.

In this study, four models were tested. The first model hypothesized that there are no groupings in the data. That is, it assumed that we were dealing with a homogenous group of test-takers. This model assumed a single latent class. However, this model is not tested in practice. It just provides a baseline model against which a model with two latent classes is evaluated. The successive models each added another latent class such that the fourth model assumed four latent classes. After 10 runs of the software (each run of the software took at least three days!), the fourth model did not converge to an acceptable solution. Hence, no fit indices for this model were provided by the software.

The fit indices for the remaining three models are reported in Table 1. Based on Clark's (2010) guidelines, the entropy index for the three-class model is adequate while the two-class model's entropy index is moderate. Similarly, the AIC index for the three-class model is lower which indicates better model fit. Both LMR and ALMR indices show that the three-class model has a significantly better fit. From among the fit indices, only BIC and its adjusted version of sample size show a better fit for the two-class

model. Based on the LMR and ALMR indices, the two-class model does not show a better fit compared to the model with no groupings. Considering all of the indices together, it is evident that the three-class model shows a better fit to the data. Hence, this model was selected for further analysis.

Table 1: Fit indices for different latent class models

Model	Ent.	AIC	BIC	SABIC	LMR	p-value	ALMR	p-value
2	0.584	210777.2	214419.3	212636.7	3060.3	0.229	3059.0	.229
3	0.745	209951.7	215418.1	212742.5	4291.6	0.000	4290.7	0.000

Figure 1 shows the performance of the three latent classes on each item. Overall, the first class has the highest performance and the third class the lowest. The university entrance exam can be regarded as both a speed and a power test. Therefore, it is natural to see a decline in the performance of the groups on items that come toward the end of the test. This is more conspicuous for the third group with a probability of almost zero on the last 25 items.

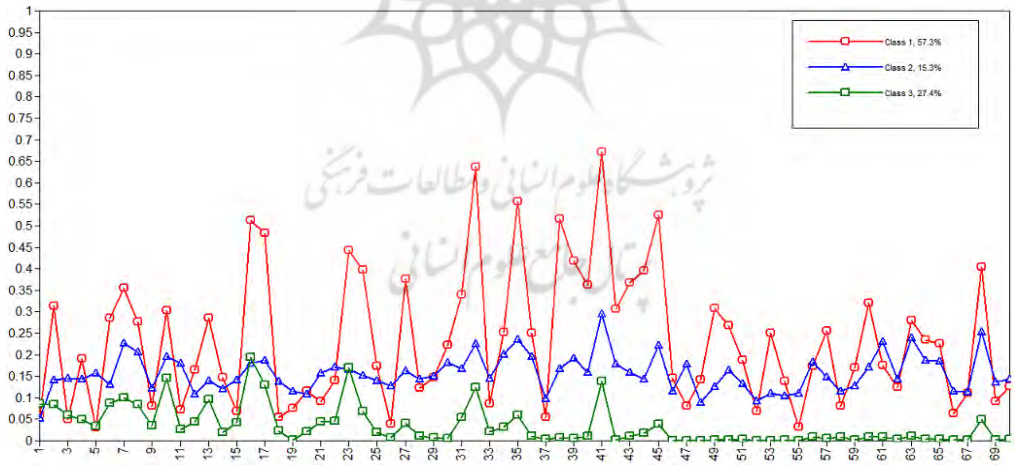

Figure 1: Performance of the latent classes on the items

Examining DIF in Latent Classes

The next phase of the study was examining DIF in the latent classes. For latent class 1, items 4, 6, 16, and 22 were flagged for uniform DIF. In addition, items 2, 4, 16, 35, 39, 42, and 50 showed non-uniform DIF.

Unlike latent class 1, a large number of items showed DIF in latent class 2. Items 6, 16, 18, 32, and 57 were flagged for significant DIF. In addition, items 2, 6, 7, 8, 9, 16, 17, 19, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 33, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 49, 52, 55, 58, 59, 61, 62, 66, 67, 69, and 70 were flagged for non-uniform DIF. It should be pointed out that whenever the main effect for the conditioning variable is not significant, even a significant effect for gender or a significant interaction would not be counted as evidence of DIF.

The results of the DIF analysis for latent class 3 revealed that no item was flagged for DIF. In addition, even the main effect for the conditioning variable was not significant for the majority of items. In other words, item responses could not be predicted even by the conditioning variable. This is cause for concern and means that we do not have evidence that examinees with higher ability had higher chances of giving a correct response to the items. The implications of this outcome for interpreting the latent class will be discussed later.

Now that the existence of DIF items in the three latent classes has been discussed, we can see if any of the covariates (i.e., high school GPA, diploma major, and quota) have a significant effect on the latent classes. In other words, the objective was to see if latent class membership can be predicted from the covariates. Since class membership (which is a nominal variable) is the dependent variable and we have multiple independent variables, this analysis can be seen as multinomial logistic regression. The results are reported in Table 2. It appears that all three covariates have a significant effect. That is, there is a statistically significant relationship between class membership (as the dependent variable) and the covariates (as the independent variables).

Table 2: Prediction of class membership by the covariates

	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept	7472.580	.000	0	.
GPA	7673.033	200.453	2	.000
Diploma	7564.377	91.797	4	.000
Quota	7585.660	113.080	4	.000

DISCUSSION

Latent class analysis showed that there were three latent classes in the sample of test-takers. This finding shows that the sample is not homogeneous. Therefore, applying any statistical analysis to this sample that assumes the sample is homogeneous would be unjustified (de Ayala & Santiago, 2017). Accordingly, DIF studies that do not take this feature of the sample into account would also provide misleading results.

The results of the present study revealed that the sample was not homogeneous. Three latent classes were extracted. A closer examination of the background variables revealed noticeable differences among the groups. While 90 percent of latent class 1 came from either Mathematics or Science, only 77 percent of latent class 2 belonged to these two majors. In addition, while about 77 of participants in latent class 1 had applied for quotas from Region 1 and Region 2, only 56 percent of participants in latent class 2 applied for these two quotas.

As for performance on the test, there are remarkable differences between the three classes. Latent class 1 had the best performance on the test and the majority of items. On the other hand, latent class 3 had by far the worst performance. Despite this variation in performance on the test, the three latent classes did not have very different high school GPAs.

An interesting outcome of the LCA analysis pertained to the results of DIF analysis for latent class 3. Remember that this class had by far the lowest performance. DIF analysis revealed that, for the vast majority of items, even the conditioning variable could not significantly predict item

response. Such a result might show that this class is the group of examinees who widely resorted to random guessing. This interpretation is in keeping with the findings of previous research which show that individuals who respond to items mostly based on random guessing can form a separate latent class (see Ulitzsch et al., 2019). Mixture models are normally suggested for modeling such data (see de Ayala & Santiago, 2017; Sen & Cohen, 2019).

The results of the present study confirmed the claim that not taking into account sample heterogeneity leads to misleading results. DIF analysis in each of the latent classes resulted in very different findings for each latent class. The initial DIF analysis for the entire sample had flagged 19 items for significant DIF. However, separate DIF analyses for each latent class did not result in similar results for any of the latent classes. Specifically, only 9 items showed significant DIF in latent class 1. On the other hand, over half of the items (i.e., 43 items) showed significant DIF in latent class 2. Unlike latent class 2 which showed so many DIF items, latent class 3 did not show even a single item with significant DIF.

The differences across the three classes were also related to their background variables. The results revealed that the background variables (i.e., Quota, GPA, and GPA) were all significantly related to the latent classes. In other words, these background variables were able to significantly predict latent class membership.

These results support the arguments in favor of an ecological approach to DIF analysis (e.g., Fox, 2003; Zumbo & Gelin, 2005). Ignoring context in DIF analysis is flawed for two reasons. First, when sample heterogeneity is not taken to account, any results from DIF analysis with the entire sample would be misleading. This is because the results obtained for the entire sample might not come true in each of the individual classes. This is exactly what the results of the present study showed.

The second reason for supporting an ecological approach to DIF analysis stems from the fact that the effectiveness of DIF analysis mostly depends on the final stage: identifying sources of DIF. DIF alone is not

evidence for bias. It can be considered as a case of bias only if the source of DIF is irrelevant to the construct being measured by the item and the test in general. In the absence of conclusive evidence, the interpretations of sources of DIF become "ad hoc" (Alavi & Karami, 2010). An ecological approach to DIF analysis overcomes this problem by including background and contextual variables in the analysis.

CONCLUSION AND IMPLICATIONS

The results of this study have serious implications for the validity and fairness of the test. The university entrance examination is clearly a high-stakes test which results have grave consequences for the test takers. Failure on the test cannot be easily ignored. This failure might be due to failure on a single item. Therefore, it is the responsibility of the test developer and user (namely, the NOET) to make sure that each and every item included in the test is fair. Our results show that this is not the case.

Apart from its implications for test validity, the difference in the number of DIF items identified in each latent class has two more implications. First, an item might be fair for a group of examinees but biased for the rest of the examinees. This is different from the common argument in the literature that if a test is biased against one group of examinees, it is biased in favor of another group. The implication of our findings offers a different conclusion. An item might be functioning fairly in one group of examinees but not so in a different group of examinees. This is usually ignored in DIF studies which are based on observed grouping variables. They have a grouping variable such as gender and inspect the performance of males and females on a given item. Since they take the entire sample of examinees as a homogeneous group, they cannot address gender DIF across subgroups of examinees.

The third latent class in this study (which had no DIF items) was the group that most probably resorted to random guessing. That is why their scores were too low and their item responses could not be predicted even

from the conditioning or matching variable. However, the other latent classes did not have this feature. They were not random guessers. However, one group had 43 DIF items and the other group had only 9 DIF items. Certainly, we cannot claim that if these two classes had an equal number of DIF items, test fairness would be ensured. This is clearly not true. However, having so many DIF items in one group might be clear evidence of test bias. This at least requires the close inspection of the test by the NOET authorities.

The final point pertains to the validation studies conducted in the ELT context. Researchers apply a wide variety of methods to validate language tests (e.g., Alavi et al., 2020; Darabad et al., 2021). Clearly, DIF analysis is also needed for a thorough examination of test validity.

Several suggestions can be offered for the follow-up research studies. This study was based on a sample of the examinees who had taken the test. Access to the entire data is not given by the NOET. However, future studies can focus on different samples from this test to see if the results of the present study are confirmed. Sampling bias can seriously affect the results.

Another fruitful avenue for future research would be applying mixture item response models to examine item parameters across latent classes. The item parameters obtained from the mixture IRT model can then be compared to item parameters obtained from a traditional IRT model. In addition, personability estimates can also be compared across the two models. The relative fit indices can be compared for these two models to see which one provides a better picture of the available data. In addition, both person ability and item difficulty estimates can be examined to see if ignoring the heterogeneity of the sample results in any bias in the estimated parameters.

Mixture models can also be used for estimating DIF (see Cohen & Bolt, 2005). Since mixture models take account of sample heterogeneity, the inspection of DIF through these models might be more dependable. In keeping with the previous paragraph, the inspection of DIF items can be done both through the mixture and traditional IRT models to see the effect

of ignoring sample heterogeneity on identifying DIF items. Presumably, the more the sample deviates from the assumption of homogeneity, the more the difference between DIF results in the mixture and traditional IRT models.

Another issue that can be examined in this direction pertains to the number and quality of background variables that entered into the LCA model as covariates. The function of these covariates is to predict latent class membership. In this study, we could include only three covariates (i.e., high school major, high school GPA, and Quota). More dependable models of test performance and its relationship to the context can be tested by including more background and contextual variables.

Another helpful line of research could focus on the mental process of test takers from different latent classes. If there is a qualitative difference across the latent classes, it might be logical to assume that they go through different mental processes in taking the test. For instance, they might resort to different test-taking strategies (Cohen, 2006). A think-aloud protocol (TAP) analysis can shed more light on the mental processes they go through and the kind of strategies they resort to.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Sayyed Mohammad Alavi http://orcid.org/0000-0003-1740-2164
Hossein Karami http://orcid.org/0000-0001-8969-3621
Mohammad Hossein Kouhpaeenejad http://orcid.org/0000-0002-3026-2955

References

Alavi, S. M., & Karami, H. (2010). Differential item functioning and ad hoc

interpretations. *TELL, 4*(1), 1-18.

Alavi, S. M., Shahsavar, M., & Norouzi, M. H. (2020). Diagnosing EFL learners' development of pragmatic competence implementing computerized dynamic assessment. *Issues in Language Teaching, 9*(1), 117-149.

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

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–280). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, S. L. (2010). *Mixture modeling with behavioral data*. (Unpublished doctoral dissertation). University of California, Los Angeles, CA.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25-29). Hillsdale, NJ: Erlbaum.

Darabad, A. M., Abbasian, G. R., Mowlaie, B., & Abusaeed, A. A. (2021). L1-Based Elicitation as a Valid Measure of L2 classroom performance assessment: Multi-method mono-trait model of validation. *Issues in Language Teaching, 10*(1), 1-36.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

de Ayala, R. J., & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology, 60*, 25-40.

Fox, J. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing, 3*(1), 21–48.

Geiser, C. (2012). *Data analysis with Mplus*. New York, NY: The Guilford Press.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.

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

Mehrazmay, R. (2012). *A DIF analysis of university entrance examination in terms of academic background. (*Unpublished MA Thesis), Al-Zahra University, Tehran, Iran.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide statistical analysis with latent variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Thousand Oaks, CA: Sage Publications, Inc.

Rindskopf, D. (2009). Latent class analysis. In R. E. Millsap & A. Mayedeu-Olivares, *The Sage handbook of quantitative methods in psychology*. London: SAGE Publications Ltd.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives, 17*, 177–191.

Shealy, R. T. & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (197–240). Hillsdale, NJ: Erlbaum.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*, 83-112.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester: John Wiley & Sons.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into the picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5*(1), 1–23.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O., & Ark, T. K. (2015). A methodology for Zumbo's third Generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*, 136-151.