ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(3); 267-285

# Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods*

Zafer ÇEPNİ ** Hülya KELECİOĞLU ***

**Abstract**

In this study, differential item functioning (DIF) and differential bundle functioning (DBF) analyses of the Academic Staff and Postgraduate Education Entrance Examination Quantitative Ability Tests were carried out. Mantel-Haenszel, logistic regression, SIBTEST, Item Response Theory-Likelihood Ratio and BILOG-MG DIF Algorithm methods were used for DIF analyses. SIBTEST was the method used for DBF analyses. Data sets for the study came from an earlier application of the examination. Gender DIF analyses showed that eleven items showed DIF. Four of the items favored male applicants, where seven of them favored female applicants. In order to investigate the sources of DIF, we consulted experts. In general, the items which could be solved using routine algorithmic operations and which are presented in the algebraic, abstract format showed DIF in favor of females. The "real-life" word problems favored males. According to DBF analyses, the operations item group favored females and the word problems item group favored males.

*Key Words:* DIF, DBF, SIBTEST, ALES

## INTRODUCTION

Large-scale tests are used to make important decisions about individuals. Large-scale exams that the Turkish community is familiar with include university entrance examinations, transition examinations for secondary education, Public Personnel Selection Examination, and Academic Personnel and Postgraduate Education Entrance Examination (Turkish acronym ALES). The first two of these exams are used for student selection. KPSS is used for staff selection and ALES is used for both student and staff selection. Over 200,000 candidates participated in ALES in 2016, which is implemented twice a year according to the information obtained from the website of the Measurement, Selection and Placement Center (Turkish acronym ÖSYM). Considering these features, ALES is one of the major large-scale exams in Turkey.

ALES consists of quantitative and verbal ability tests. Quantitative ability tests aim to measure quantitative and logical reasoning skills. The tests include items that candidates who have graduated from different bachelor's programs can answer correctly (ÖSYM, 2008). When the content of the ALES quantitative tests used in different years is examined, it is observed that the subject areas of the materials, in general, do not exceed the ninth-grade level. Content areas like trigonometry, complex numbers, limit, derivatives and integrals with which only the students in quantitative branches of high schools would be familiar are not included in the ALES quantitative tests. The difference between the quantitative 1 test and the quantitative 2 test is described as "more advanced items are used in the quantitative 2 test" (ÖSYM, 2008).

It is an indispensable requirement to present validity evidence for the large-scale examinations in which important decisions are made about candidates. One of the major threats to efficacy is item and test bias (Clauser & Mazor, 1998). For this reason, the scores should be fair to different groups taking the exams. Test fairness is not only a technical issue within the validation procedure but also an issue having

* This paper was derived from the first author's doctoral thesis titled *Detecting differential item functioning using SIBTEST, Mantel-Haenszel, logistic regression and Item Response Theory methods*.

** Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, cepni@hacettepe.edu.tr, ORCID ID: 0000-0002-8033-905X

*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
Çepni, Z. & Kelecioğlu, H. (2021). Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods. *Journal of Measurement and Evaluation in Education and Psychology, 12*(3), 267-285. doi: 10.21031/epod.988879
*Received: 31.08.2021*
*Accepted: 24.09.2021*

political, philosophical, economic, social and legal aspects (Camilli, 2006). In this framework, providing empirical evidence for test fairness is considered an important part of test development and validity studies (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014; Joint Committee on Testing Practices, 2004).

In order to understand test fairness, the concepts of item effect, statistical bias and differential item functioning (DIF) should be explained. Different performances of different groups on an item or a test is called item effect or test effect. Observation of the item or test effect does not necessarily mean that the item or test is biased (Clauser & Mazor, 1998; Millsap & Everson, 1993). If the cause for the different performances is seen as the item itself, then there is statistical bias. Here, there are differences according to groups in estimating some parameters. Statistical bias could appear in two ways. Firstly, the item parameters in the measurement model may be different for groups. This can be explained as DIF in the sense that impairment of measurement equivalency with regard to internal criteria. In the analysis of this kind of situation, an answer to the question “Does this item measure the same variable that the rest of the exam measures?” is sought. Secondly, the intercept or slope of the line used in predicting an external criterion or the standard error of the prediction may differ for different groups. This situation could be expressed as impairment of measurement equivalency with regard to external criteria or differential prediction. In an analysis of this kind of situation, the question is whether the test measures the same construct for both groups according to an external criterion (Camilli, 2006).

DIF refers to the fact that the performances of individuals from the reference and focus groups at the same level of ability are different. An item that does not exhibit DIF has the same measurement properties for reference and focus groups. In other words, for an item that does not show DIF, the likelihood of individuals with equal ability to respond to the item correctly is the same even if the individuals belong to different groups. However, if different item difficulties are observed in different groups of equal skill levels, the item exhibits DIF (Millsap & Everson, 1993). Since the DIF analyses are based on internal criteria, they assume that other validity evidence is sufficient (Clauser & Mazor, 1998). Therefore, it is generally appropriate to establish the factor structure of the tests before DIF analyses.

Although tests are often considered unidimensional, it is rare that the ability to answer an item correctly is only one. Within the multidimensionality-based DIF paradigm framework, the groups are statistically matched on the primary factor measured by the test, θ. The secondary skills required to correctly answer the item in the same paradigm are considered as η. If the groups differ on the secondary skills that the items measure, DIF is seen in these items. In other words, the reason why the item shows DIF is the difference between the groups on the secondary factor (η). There is a secondary variable (η) that is effectively functioning in a DIF item. This secondary variable, which leads to DIF, can be determined by examining the item by experts. The decision of flagging the item as biased or not is based on what the secondary variable is. If the experts see the secondary variable as an element not to be included in the construct measured by the test, the item is labelled as biased and should be removed from the test. For example, if a secondary variable such as "familiarity with hunting terms" plays a role in the analysis of any material in the test of reading skills, it may be suggested to remove the item from the test (Ackerman, 1992). If these secondary variables are deemed as integral to the construct being measured, then the item is not considered biased—only a DIF item. For example, word problems in mathematics tests may show DIF because of the effect of reading skills in their responses. Whether this DIF should be taken as bias is determined by assessing whether the reading skill is a secondary variable considered to be measured by these items. If the reading skills are a secondary variable that is desired to be measured by those items, the items are treated as DIF items only, not biased. If the undesired variables lead to DIF, the item could be considered as biased (Zumbo & Gelin, 2005). In this framework, DIF is a necessary but not sufficient condition for items to be biased (Zumbo, 1999).

As Ong, Williams and Lamprianou (2011) point out, the bias decision depends on the boundaries of the target construct to be measured, and clear cut limits are not always published or easy to draw. For example, when algorithmic procedural knowledge items function in favour of female candidates, it seems that the ability to perform operations in a step-by-step and organized manner is also effective in these items as well as general quantitative skills. Significant differences in the secondary construct between male and female candidates lead to DIF in such items. Whether or not these items will be flagged as biased would be determined by whether the ability to perform those procedures in a step-by-step and organized manner is within the target construct to be measured.

The main purpose is to eliminate item bias, which is an important threat to test validity. In this case, it is advisable to remove the item from the test if the part causing the bias in the item cannot be corrected (Ackerman, 1992, Camilli, 2006, Clauser & Mazor, 1998). Determining and eliminating item bias is used for improving test validity. In this framework, it is especially important to determine the causes of DIF as well as to detect DIF items.

Potentially biased items are detected using DIF methodology. The aim here is to identify and eliminate bias resulting from test design, content and item types among different gender, ethnicity, language, culture groups and ultimately to increase test validity (AERA, *et al*. 2014). Since potentially biased items are determined using DIF analyses, DIF detection could be considered as a step in item bias detection.

Once DIF items are identified, the variables that are the source of DIF should be examined for the decision to flag the items as biased or not (Clauser & Mazor, 1998). In DIF analyses, grouping variables can be gender, country, culture, language, socioeconomic level or ethnicity (Camili, 2006). Important DIF sources in the cross-cultural assessments which are used for international comparisons are translation inadequacies, lack of the same reciprocal of concepts in different cultures, different levels of familiarity with different concepts from different cultures, different curricula of different countries and different teaching methods and qualifications that are emphasized by different curricula (Asil, 2010; Ercikan, 1998; Grisay, de Jong, Gebhardt, Berenzer, & Halleux-Monseur, 2007; Hambleton, Merenda, & Spielberger, 2005; Yıldırım and Berberoglu, 2009). Factors like item format, content and cognitive complexity level are among the popular gender DIF sources (Bakan Kalaycığolu & Berberoğlu, 2010; Bakan Kalaycıoğlu & Kelecioğlu, 2011; Mendes-Barnett & Ercikan, 2006; Zumbo & Gelin, 2005).

If a DIF item is functioning in favor of a group at all levels of ability, this is called uniform DIF. The item characteristic curves determined for the two groups of such an item do not intersect. An item with intersecting characteristic curves tends to favor a group to a certain level of ability and favors the other group at higher levels of skill. This is called a non-uniform DIF (Hambleton, Swaminathan, & Rogers, 1991). Only uniform DIF items are investigated in this research because uniform DIF items favor one group more significantly and the interpretations of non-uniform DIF are more complicated (Smith & Reise, 1998).

### *DIF Detection Methods*

In DIF determination methods, the individuals in the two groups, which are generally taken as focus and reference, are matched according to their ability estimation. For these matched groups, DIF statistics are calculated using the correct response rates of the items. A hypothesis is constructed regarding the item, saying that the item functions equivalently between the groups, and a statistical significance test is performed. However, statistical significance tests are not considered satisfactory for the interpretation of the practical significance and effect size of DIF (Camilli, 2006). Therefore, methods that provide effect size statistics may be more useful in practice. Although the methods generally give similar results to some extent, they are not in perfect agreement because they use different algorithms and different matching criteria. In addition, the cut-off points they use to flag the DIF items are different (Bakan Kalaycıoğlu & Berberoğlu, 2010; Doğan & Öğretmen, 2008; Gök, Kelecioğlu & Doğan, 2010). For this reason, it is recommended that researchers and test developers use multiple methods for DIF analysis (Hambleton, 2006).

It is possible to divide DIF detection methods into two groups as (1) methods using the observed raw scores in matching of individuals and (2) methods based on Item Response Theory (IRT) (Camilli, 2006). Mantel-Haenszel (Holland & Thayer, 1988), logistics regression (Swamanithan & Rogers, 1990) and SIBTEST (Roussos & Stout, 1996a) are among the former group. Restricted factor analysis is another method based on factor analysis that does not lend itself in either group (Oort, 1992).

### *Mantel-Haenszel*

The Mantel-Haenszel (MH) is a DIF detection method given by Holland and Thayer (1998) in the measurement literature. In this method, the total test score is used as a matching criterion. The total test score is treated as a discrete variable in constructing the equivalent ability examinees for focus and reference groups.

For analysis, a three-dimensional matrix of size *2 × 2 × S* is formed, where *S* is the number of ability levels being generated according to the correct and incorrect answers of the individuals from different groups. For each ability level of focus and reference groups, a data structure as shown in Table 1 is analyzed.

Table 1. Data Structure Used in Mantel-Haenszel

| Group | Correct | Incorrect | Total |
|---|---|---|---|
| Reference | $A_j$ | $B_j$ | $n_{Rj}$ |
| Focus | $C_j$ | $D_j$ | $n_{oj}$ |
| Total | $m_{1j}$ | $m_{oj}$ | $T_j$ |

A likelihood ratio is obtained by using the values in the tables for each ability level. This ratio is given in Equation 1.

$$\alpha_{MH} = \frac{\sum_j A_j \, D_j / T_j}{\sum_j B_j \, C_j / T_j} \qquad (1)$$

The final output of the Mantel-Haenszel algorithm is the $\Delta_{MH}$ statistic, which is -2.35 times the natural logarithm of this likelihood ratio. Since the standard error of this statistic is known, a hypothesis test can be performed using a $\chi^2$ distribution. Negative values of the $\Delta_{MH}$ statistics indicate that the item is in favor of the reference group and the positive values indicate that the item functions in favor of the focus group. In addition, since $\Delta_{MH}$ is itself an effects size measure, it can be used to interpret the practical significance of DIF. A commonly used categorization schema has been proposed by Zieky (1993), which is shown in Table 2. Mantel-Haenszel DIF statistics can be calculated by means of EZDIF software (Waller, 1998).

Table 2. Interpretation of Mantel-Haenzsel DIF Statistic

| Level | Value | DIF amount |
|---|---|---|
| A | $\lvert\Delta_{MH}\rvert < 1$ | None or negligible |
| B | $1 \leq \lvert\Delta_{MH}\rvert < 1.5$ | Middle |
| C | $\lvert\Delta_{MH}\rvert \geq 1.5$ | High |

### *Logistics regression*

Swamanithan and Rogers (1990) have shown that logistic regression (LR) can be used to detect DIF. In this method, the matching criterion is the total test score. However, unlike the Mantel-Haenzsel method, it is taken as a continuous variable. Group affiliation and total test score are independent variables in the logistic regression, whereas the response to the item is a dependent variable. The mean for different groups of an item is expressed in Equation 2 in the expected value.

$$\varepsilon(Y_i \mid X_i, G_i) = P_i \qquad (2)$$

The LR equation for uniform DIF is constructed as shown in Equation 3.

$$Z_i = ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_i + \beta_2 G_i \tag{3}$$

An interaction term is added to the regression equation for non-uniform DIF analysis. For a hypothesis test of whether the item being inspected exhibits uniform DIF, the fit of the above model and the fit of the model obtained by subtracting the group variable can be compared. The difference between the R$^2$ values of the two models, ΔR$^2$, indicates an effect size used to interpret the amount of DIF (Zumbo, 1999). For the interpretation of ΔR$^2$ values, Zumbo and Thomas (1996) and Jodoin and Gierl (2001) proposed two separate classifications given in Table 3.

Table 3. Recommended Categories of Classification for Interpreting the ΔR$^2$ Values

| Level | Zumbo and Thomas (1996) | Jodoin and Gierl (2001) | DIF amount |
|---|---|---|---|
| A | $\Delta R^2 < .13$ | $\Delta R^2 < .035$ | None or negligible |
| B | $.13 \leq \Delta R^2 < .26$ | $.035 \leq \Delta R^2 < .070$ | Middle |
| C | $\Delta R^2 \geq .26$ | $\Delta R^2 \geq .070$ | High |

DIF statistics calculated by Mantel-Haenzsel and logistic regression methods are quite consistent when the index values are considered, but regarding the cut-off points used in the categoricals this consistency seems to be inadequate (Bakan Kalaycıoğlu & Berberoğlu, 2010, Doğan & Öğretmen, 2008; Gök, vd. 2010). In addition, Higaldo and Lopez-Pina (2004) tested the effectiveness of logistic regression and some other methods to detect DIF under simulation conditions and showed that only 1% of the DIF items were flagged when the cut-off point .13 is used, and 20% when .035 used. As a result of this study, it was emphasized that new criteria should be determined for the interpretation of ΔR$^2$ statistic. Due to this condition of the ΔR$^2$ statistic, Bakan Kalaycıoğlu and Kelecioğlu (2011) used the first cut-off point as .010 and the second as .020, taking into account the ΔMH DMF index. Logistic regression DIF analysis can be performed in SPSS software using the SPSS codes provided by Zumbo (1999).

### *SIBTEST*

SIBTEST method, developed by Shealy and Stout (1993), can be used in determining statistically whether or not one item and more than one item displays DIF. The item or items for which DIF analysis is to be performed is/are included in a group and the other items are put in another group and thus, the test is divided into two parts. Matching is done with the actual scores estimated by means of the total scores on the items in the second group, and the performance of the groups which are analysed for DIF is compared (Gierl, 2005). The expected scores of the applicants in the reference (R) and focus (F) groups are identified in Equations 4 and 5- where *k* is the score received from DIF item or items, $P_{Rk}(t)$ and $P_{Rk}(t)$ are the ratios of t score and the applicants receiving the *k* scores on the items.

$$ES_R(t) = \sum_k kP_{Rk}\,(t) \tag{4}$$

$$ES_F(t) = \sum_k kP_{Fk}\,(t) \tag{5}$$

These two values are used by correcting for measuring errors in the SIBTEST. In this case, the final output of the SIBTEST method, $\beta_u$ DMF index, is derived as in Equation 6.

$$\beta_u = \sum_t \left( \left[ES_R\,(t) - ES_F\,(t)\right] \left[\frac{N_R(t) - N_F(t)}{N}\right] \right) \tag{6}$$

$N_R(t)$ and $N_F(t)$ values in the formula indicate the number of applicants whose matching scores are t in the reference and focus groups. Because the standard error of $\beta_u$ index is known, a result of a hypothesis test can be obtained. $\beta_u$ index indicates an effect size. The classification developed by Rousssos and Stout (1996b) to interpret the amount of DIF is shown in Table 4. SIBTEST can be performed by using the software called SIBTEST (Stout & Roussos, 1995).

Table 4. Classification Categories Recommended for the Interpretation of $\beta_u$ Values

| Groups | Values | Amount of DIF |
|---|---|---|
| A | $\beta_u < 0.059$ | None or negligible |
| B | $0.059 \leq \beta_u < 0.088$ | Middle |
| C | $\beta_u \geq 0.088$ | High |

SIBTEST can also test whether or not more than one item display DIF synchronically. In the same vein, $\beta_u$ index also shows the amount of DIF for more than one item. Yet, no systems of classification were recommended for the evaluation of the amount of DIF when used for more than one item (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Bisanz, Bisanz, & Boughton, 2003; Ong, et al. 2011). It is possible to relatively compare the $\beta_u$ statistics of both groups of items. SIBTEST is a technique based on the fact that the skills necessary for responding to an item correctly are multidimensional. In this framework, when the primary skill necessary for responding to an item correctly is taken as $\theta$ and the secondary skill as $\eta$, differentiation of the distribution of different groups on $\eta$ is considered to be the source of DIF (Roussos & Stout, 1996a). SIBTEST can be used in determining the characteristics of items displaying DIF, in testing the DIF hypotheses which can be constructed beforehand and in making healthier generalisations about the sources of DIF due to the fact that SIBTEST enables one to group items and to perform DIF analysis on them (Gierl, et al. 2003; Mendes-Barnett & Ercikan, 2006).

### *Item Response Theory-Likelihood Ratio*

As the name suggests, the item response theory likelihood ratio (IRT-LR) is an IRT-based method (Thissen, Steinberg, & Wainer, 1993). Therefore, IRT-based ability estimations, and not observed scores, are used in matching individuals. The IRT-LR analyses can be performed on IRTLRDIF software (Thissen, 2001). First, a generalised model in which item parameters are freed for both groups is constructed in DIF analysis in which IRT-LR is performed. After that, the restricted model enabling one to restrict the item parameters in the same way for both groups is constructed. -2log likelihood ratios are compared for the fit between the two models. The difference between the two models is reported as $G^2$ statistics.

$G^2$ statistics makes it possible to perform a synchronic hypothesis test about whether or not all the parameters are equal in the two groups. The $G^2$ value is compared with the critical value of $\chi^2$ distribution, which is the number of parameters in the degrees of freedom IRT model, and thus the hypothesis is tested. If the synchronic hypothesis testing is found to be significant for all parameters, the $G^2$ value is compared with 3.84- which is the critical value of single freedom degree $\chi^2$ distribution- for difficulty and discrimination parameters and thus, hypotheses are tested. When the $G^2$ value used in synchronic parameter comparisons is below 3.84, it is impossible for any parameters to be algebraically significant. For this reason, the IRTLRDIF software cannot perform the test for individual parameters in such cases (Thissen, 2001). $G^2$ is not an effect size statistics. It is recommended that anchor items be selected by considering the other initial IRT-LR analysis and the other DIF statistics be used in IRT-LR analyses (Wang & Yeh, 2003). Six anchor items were selected for each IRT-LR analysis in this study. The $G^2$ values derived from the initial application of IRT-LR method and the other DIF statistics were taken into consideration in selecting the anchor items.

### *BILOG-MG DMF Algorithm*

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) software offers an IRT-based algorithm for DIF analyses. In this algorithm, parameters are estimated for two separate groups in a way similar to unequal groups test matching design, and they are brought on the same scale. The difference of difficulty parameters that are brought on the same scale and the standard errors for the difference are reported (du Toit, 2003). A hypothesis test is done by dividing adjusted difficulty difference values ($\Delta$b) into standard errors. The $\Delta$b values express the effect size about the magnitude of DIF amount (Smith & Reise, 1998). Yet, there are no widely used classifications of these values. BILOG-MG algorithm allows item discrimination to differ from item to item, but it does not allow differences between groups. Therefore, it is appropriate for use only in determining and interpreting uniform DIF (Smith & Reise, 1998). It is necessary to show that IRT assumptions are satisfied prior to IRT-based DIF analyses. Therefore, unidimensionality was tested in this study prior to DIF analyses.

Although DIF analyses yield consistent results on considering the indices, it is observed that they do not determine the same items as DIF display in items on considering the cut-off points (Higaldo & Lopez-Pina, 2004). Thus, it is recommended to use more than one method in DIF analyses (Hambleton, 2006). In line with this recommendation, more than one method was used in this study to detect DIF.

### *Differential Bundle Functioning*

Items that are probable to be biased are detected through DIF analyses. However, the causes of different functioning in different groups cannot be detected through DIF analyses. Differential bundle functioning (DBF) analyses can be used in determining the sources of DIF, and thus it becomes possible to analyse whether or not the sources of DIF are accepted into the construct intended to be measured (Ong et al., 2011). These analyses test whether or not items having certain properties function as a group. In some cases, the amount of DIF displayed by items is lower than B or C levels; but when such items come together, the effect of the item group is more remarkable and it should be taken into account (Nandakumar, 1993). DBF analysis is appropriate for analysing such situations.

DBF analyses can be performed in SIBTEST method (Roussos & Stout, 1996). In this method, item groups are formed according to their certain properties and whether or not the item groups function in different ways for different groups of students is analysed. In consequence of DBF analyses, which can be used on SIBTEST software, $\beta_u$ DBF statistics are calculated for each group of items. A hypothesis test is done with the significance level of these statistics. For item groups, $\beta_u$ statistics is an effect size statistics expressing the amount of DBF. But no widely used schema is available for item groups level classification (Gierl et al., 2001; Gierl, et al. 2003; Ong, et al. 2011). There are studies trying to determine the sources of DIF by doing DBF analysis on pre-determined item groups in the literature (Gierl, et al. 2001; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001). Shedding light on the sources of DIF in addition to determining DIF displaying items is considered as a component of finding validity evidence (Ong, et al. 2011).

### *Purpose*

This study aims to determine the items displaying DIF as well as item groups displaying DBF in ALES quantitative ability tests according to gender and to compare the results of differing DIF detection methods in a real data set. DIF and DBF analyses were performed for this purpose in ALES quantitative ability tests administered in Fall 2008. In this way, the target was to determine the items displaying DIF in ALES quantitative ability test and to reveal the causes for different functioning of items according to groups by using DBF analyses.

## METHOD

### *Data Set*

The raw data necessary for the study were obtained from ÖSYM. After obtaining the entire national data set from the application of ALES, the candidates who responded to at least one item correctly were taken into consideration. The analyses were carried on the population so as to prevent the errors from being caused by sample formation. Yet, SIBTEST software can work with data sets having 7,000 participants in each group. Therefore, samples of randomly selected 13,000 applicants from quantitative test 1 and 11,000 applicants from quantitative test 2 were formed for analyses to be performed through SIBTEST by using SPSS software. The whole data set was used in data analyses apart from SIBTEST. The whole data set for quantitative test 2 and the distribution of the sample according to gender and department scores are shown in Table 5. The data set included 133,788 applicants for quantitative test 1 and 103,088 applicants for quantitative test 2. Of the applicants, 51% were female, whereas 49% were male in the quantitative test 1. The proportion was also similar in quantitative test 2. It was found that the data set chosen for SIBTEST sampling represented the data set taken as the population in terms of such variables as gender and department.

Table 5. Distributions of Scores

| Gender | Quantitative 1 Test | | | | Quantitative 2 Test | | | |
|---|---|---|---|---|---|---|---|---|
| | Whole data set | | SIBTEST sample | | Whole data set | | SIBTEST sample | |
| | *N* | *%* | *n* | *%* | *N* | *%* | *n* | *%* |
| Female | 68170 | 51 | 6629 | 51 | 53725 | 52 | 5636 | 51 |
| Male | 65618 | 49 | 6371 | 49 | 49363 | 48 | 5364 | 49 |
| Total | 133788 | 100 | 13000 | 100 | 103088 | 100 | 11000 | 100 |

### *Data Analysis*

The data were coded by marking corrects answers as 1 and marking incorrect or empty answers as 0. Prior to DIF analyses, a unidimensional measurement model was tested through confirmatory factor analysis by means of the asymptotic covariance matrix for quantitative 1 and quantitative 2 tests in order to test the unidimensionality of the data coming from the tests. PRELIS software was used in deriving asymptotic covariance matrix, whereas SIMPLIS software was used in performing the confirmatory factor analysis. Score distributions for the tests were determined and the test statistics and α coefficients were calculated. In addition to that, the item difficulties and discrimination indices for the overall test and for the sub-groups were also calculated.

Mantel-Haenszel, logistic regression, SIBTEST, IRT-LR and BILOG-MG DIF algorithm techniques were used in determining the items displaying DIF. Mantel-Haenszel analysis was done by using EZDIF software, logistic regression analysis was performed by using the codes provided by Zumbo (1999) and by using SPSS software, SIBTEST was performed by using the software carrying the same name, IRT-LR analysis was performed by using the software IRTDIF and BILOG-MG DIF analysis was performed by using the software carrying the same name. It was found that the results of almost all hypothesis tests performed with logistic regression, IRT-LR and BILOG-MG were significant. Due to the fact that the data set used was very large, the items were marked as at least middle (B level) according to at least two methods according to the classification of the effect size of Mantel-Haenszel, logistic regression and SIBTEST techniques were determined as items displaying DIF. The $G^2$ statistics provided by IRT-LR for the items whose indices were calculated to be very close to the cut-off scores used in the classification and the Δb statistics provided by BILOG-MG were also taken into consideration. Since there was not a schema for classifying these two techniques, the evaluation was made by comparing the other items displaying relative DIF.

Expert opinion was consulted for the causes of different functioning of DIF displaying items. Four of the eight experts included in the study held Ph.D. in measurement and evaluation while one had a doctorate degree in science education, one had a doctorate in mathematics education, one was a student of the doctorate in measurement and evaluation and one was a student of the doctorate in mathematics education. The items displaying DIF and the directions in which they displayed DIF were shown to the experts, and their opinions on the causes for the items to display DIF were obtained via open-ended questions. The forms in which the experts stated their opinions were sent through e-mails, and the experts were also interviewed face to face. Relevant literature, as well as DIF results, was taken into consideration in DBF analyses and bundles of items having certain properties were formed. DBF analyses were conducted with the bundles of items by using SIBTEST.

### *Unidimensionality, test and item statistics*

A unidimensional measurement model was tested with confirmatory factor analysis through asymptotic variance matrix for each of Quantitative 1 and Quantitative 2 tests so as to test the unidimensionality of the data coming from the tests. Consequently, the unidimensional measurement model was found to have an adequate model-data fit for both tests. The model-data fit statistics for the factor analysis are shown in Table 6 and the descriptive statistics for Quantitative 1 and Quantitative 2 tests are shown in Table 7. A close examination of the statistics in Table 7 demonstrates that male participants have a slightly higher average than female participants but that they have similar score heterogeneity. It is also clear that the tests slightly differ in average discrimination according to gender groups.

Item discriminant indices for Quantitative 1 and Quantitative 2 tests took on values in the 0.40-0.93 and 0.43-0.92 range. On the other hand, ALES is an examination in which a correction formula is used against accidental success, and it is thought that applicants rarely give incidental answers to the test items. Thus, an accidental success parameter was not needed in modelling the data. Therefore, a 2-parameter logistic model was chosen in the analyses of BILOG-MG DIF algorithm and in IRT-LR analyises- which were IRT-based analyses. The scatter diagrams for the item statistics of Quantitative 1 and Quantitative 2 tests are shown in Figure 1. An examination of data concerning item difficulty makes it clear that the items in the tests rank from the easiest to the most difficult in their difficulty in a wide range. It may be stated that the items in the tests generally have high discriminating power, considering item discrimination.

Table 6. Fit Indices for Confirmatory Factor Analysis

| Indices | Quantitative 1 | Quantitative 2 |
|---|---|---|
| SB$\chi^2$ | 278754.77 | 342093.08 |
| Degrees of freedom | 740 | 740 |
| RMSEA | 0.053 | 0.067 |
| SRMR | 0.058 | 0.062 |
| NFI | 0.99 | 0.99 |
| CFI | 0.99 | 0.99 |

Table 7. Test Statistics

| Statistics | Quant. 1 | Quant. 2 | Quant. 1 | | Quant. 2 | |
|---|---|---|---|---|---|---|
| | Overall | | Male | Female | Male | Female |
| Number of applicants | 133788 | 103088 | 68170 | 65618 | 53725 | 49363 |
| Mean | 24.19 | 23.54 | 25.14 | 23.20 | 24.39 | 22.52 |
| Standad deviation | 11.3 | 11.4 | 11.43 | 11.09 | 11.59 | 11.04 |
| Skewness | -0.48 | -0.43 | -0.57 | -0.40 | -0.54 | -0.33 |
| Kurtosis | -1.04 | -1.10 | -0.97 | -1.07 | -1.03 | -1.14 |
| Average difficulty | 0.60 | 0.59 | 0.63 | 0.58 | 0.61 | 0.57 |
| Average discrimination | 0.75 | 0.76 | 0.77 | 0.73 | 0.78 | 0.73 |

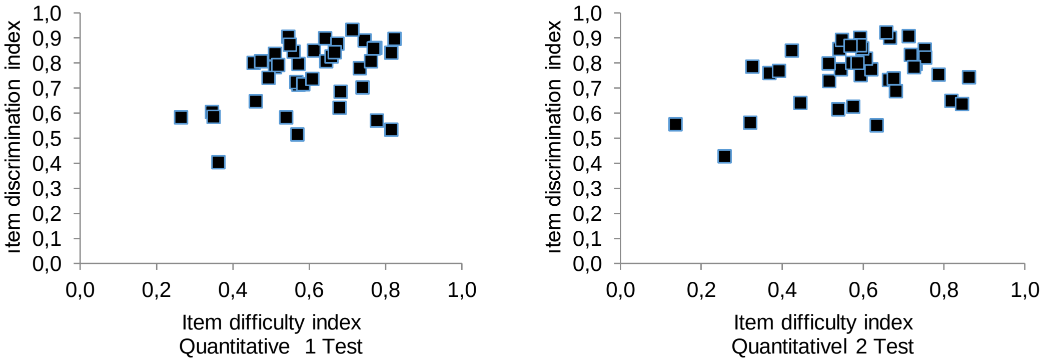


Figure 1. Scatter Diagrams for Item Statistics in Quantitative 1 and Quantitative 2 Tests

It was found that the difference between gender groups in Quantitative 1 test item difficulty was 0.18 at the maximum and the difference in item discrimination was 0.08 at the maximum. For the Quantitative 2 test, on the other hand, the difference in item difficulty was found to be 0.12 at the maximum and the difference in item discrimination was found to be 0.14 at the maximum. On estimating the item parameters within the framework of IRT separately according to applicant groups, they were not available on the same scale. Availability of item parameters in the framework of IRT on the same scale for the groups is made possible in IRT-based DIF analyses. Therefore, item statistics within the framework of IRT are given in relevant DIF analyses.

## RESULTS

### *Findings for DIF Analyses*

DIF analyses on Quantitative 1 test were performed in Mantel-Haenszel, logistic regression, SIBTEST, IRT-LR and BLOG-MG methods. The statistics considered in determining the DIF displaying items were as in the following: $\Delta_{MH}$ for MH, $\Delta R^2$ for LR, $\beta_u$ for SIBTEST, $G^2$ for IRT-LR and $\Delta b$ for BILOG-MG DIF algorithm. Findings on DIF obtained through MH, LR and SIBTEST- which are the methods based on observed scores- are shown in Table 8. This study takes the values of 1 and 1.5 for MH, 0.010 and 0.020 for LR and 0.059 and 0.088 for SIBTEST as the criteria in determining DIF levels. The tables include only DIF displaying items. The findings for IRT-LR and BILOG DIF algorithm, which are IRT-based DIF analyses, are shown in Table 9. IRTLRDIF software uses anchor items in bringing item parameters onto the same scale. Anchor items were determined by taking other DIF statistics and item difficulty levels into consideration in this study.

Table 8. Findings for MH, LR and SIBTEST

| Item no | MH | | LR | | SIBTEST | | Advantaged group |
|---|---|---|---|---|---|---|---|
| | $\Delta_{MH}$ | Level | $\Delta R^2$ | Level | $\beta_u$ | Level | |
| Quantitative 1 | | | | | | | |
| 1 | -0.975 | | 0.011 | B | 0.081 | B | Male |
| 6 | 1.383 | B | 0.006 | | -0.049 | | |
| 9 | -0.981 | | 0.007 | | 0.060 | B | |
| 10 | 1.998 | C | 0.014 | B | -0.077 | B | Female |
| 11 | 1.211 | B | 0.006 | | -0.037 | | |
| 16 | 1.587 | C | 0.011 | B | -0.085 | B | Female |
| 17 | 1.847 | C | 0.015 | B | -0.113 | C | Female |
| 18 | 1.039 | B | 0.003 | | -0.056 | | |
| 20 | 1.436 | B | 0.009 | | -0.082 | B | Female |
| 21 | -1.751 | C | 0.019 | B | 0.115 | C | Male |
| 30 | -1.024 | B | 0.015 | B | 0.101 | C | Male |
| Quantitative 2 | | | | | | | |
| 5 | 1.397 | B | 0.012 | B | -0.061 | B | Female |
| 15 | 1.709 | C | 0.014 | B | -0.085 | B | Female |
| 16 | 1.391 | B | 0.009 | | -0.073 | B | Female |
| 23 | -1.224 | B | 0.006 | | 0.065 | B | Male |
| 36 | -0.915 | | 0.004 | | 0.059 | B | |

On considering the DIF statistics, which were obtained from IRT-based methods, it was found that almost all the items were marked as DIF displaying items. Therefore, the items which were marked as items having DIF according to at least two of the methods of MH, LR and SIBTEST were considered as DIF displaying items; and which groups they offered advantages was analysed. Accordingly, items 1, 21 and 30 in Quantitative 1 test and item 23 in Quantitative 2 test functioned in favour of male applicants while items 10, 16, 17 and 20 in Quantitative 1test and items 5, 15 and 16 functioned in favour of female applicants.

Table 9. Findings for IRT-LR and BILOG-MG

| Item no | MTK-OO | | | | | BILOG-MG | | | | Advantaged group |
|---|---|---|---|---|---|---|---|---|---|---|
| | $G^2$ | $A_{male}$ | $A_{female}$ | $B_{male}$ | $B_{female}$ | a | $B_{male}$ | $B_{female}$ | $\Delta b$ | |
| Quantitative 1 | | | | | | | | | | |
| 1 | 1157.1 | 1.16 | 1.11 | -0.25 | 0.13 | 0.72 | -0.44 | -0.09 | 0.35 | Male |
| 10 | 1923.4 | 2.41 | 2.27 | -0.58 | -0.94 | 1.58 | -0.72 | -1.07 | -0.35 | Female |
| 16 | 1632.3 | 2.09 | 2.02 | -0.29 | -0.62 | 1.34 | -0.44 | -0.77 | -0.33 | Female |
| 17 | 2066.5 | 2.51 | 2.54 | 0.33 | 0.02 | 1.53 | 0.17 | -0.16 | -0.33 | Female |
| 20 | 1378.0 | 2.11 | 2.38 | 0.20 | -0.07 | 1.31 | 0.04 | -0.26 | -0.29 | Female |
| 21 | 2514.3 | 2.21 | 2.41 | -0.13 | 0.23 | 1.29 | -0.32 | 0.03 | 0.35 | Male |
| 30 | 1384.9 | 0.80 | 0.68 | 0.65 | 1.40 | 0.46 | 0.42 | 1.02 | 0.60 | Male |
| Quantitative 2 | | | | | | | | | | |
| 5 | 845.9 | 1.43 | 1.22 | -1.18 | -1.77 | 0.83 | -1.33 | -1.74 | -0.41 | Female |
| 15 | 327.4 | 1.93 | 1.84 | -0.58 | -0.97 | 1.14 | -0.72 | -1.07 | -0.36 | Female |
| 16 | 807.9 | 2.12 | 1.97 | -0.54 | -0.85 | 1.24 | -0.69 | -0.95 | -0.26 | Female |
| 23 | 527.5 | 2.93 | 2.78 | -0.49 | -0.33 | 1.57 | -0.69 | -0.47 | 0.21 | Male |

Item 1, which functioned in favour of male applicants in Quantitative 1 test, required skills related to ordering fractions. The item was presented in a way that takes too much time to solve in algorithmic methods such as equalising denominators. Therefore, applicants needed to imagine how behind the fraction is on the line of numbers according to 1so that they could solve the problem given in the item. In this aspect, it was found that the item differed from abstract items, which could be solved in algorithmic operations. Real-life situations were presented verbally or in the form of tables in the other three items, which displayed DIF in favour of male applicants. The applicants were required to solve the problem which was developed through real-life situations by using mathematical reasoning skills. It was apparent that the three problems involved cognitive processes more complex than algorithmic operation skills. One of those items is shown in Figure 2 below. We provide English translations of the items here. The original items in Turkish could be found in the Turkish version of this paper and in fulltext of the first author's doctoral dissertation.

In the table below, the number of people immigrating to other countries from countries A, B, C, D and E with a certain population in 2007, the number of people immigrating to these countries from foreign countries, the number of people born and died in these countries are given.

| | COUNTRIES | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | E |
| Immigrating from | 1600 | 4200 | 5000 | 4800 | 3400 |
| Immingrating to | 5400 | 4800 | 7000 | 1000 | 3800 |
| Born | 3200 | 5800 | 1300 | 3400 | 5200 |
| Died | 2000 | 3400 | 3300 | 3600 | 2600 |

I. The population of country C has not changed.
II. At the end of the year, the population of country B is equal to the population of country E.
III. There are two countries with declining population.

Given the information in the table, which of the above are definitely true?

A) Only I B) Only II C) I and II
D) I and III E) II and III

Figure 2. An Item Displaying DIF in favour of Male Applicants

It was found that six items displaying DIF in favour of female applicants were the items which could be solved with algorithmic operations given in abstract algebraic expressions. A sample for such an item functioning in favour of female applicants is shown in Figure 3. It was also clear that another example, item 20 included in Figure 3 and which also functioned in favour of female applicants, was also a real-life problem and that it was also an item using vehicles and the concept of speed as the context. The fact that the item functioned in favour of female applicants was an unexpected situation. Almost all of the experts included in the research stated that they had expected the item to function in favour of male applicants. Yet, one of the experts said that the item was expected to function in favour of male applicants but the fact that the problem could be solved by using the equation Distance=speed X time directly and that the proportions of the distance covered by the two cars or the differences could not be used might have caused the item to function in favour of female applicants rather than male applicants. It was found in studies that real-life problems of this type functioned in favour of male applicants (Harris & Carlton, 1983; Mendes-Barnet & Ercikan, 2006). Whether or not this situation observed in ALES examinations which were in contrast to the case in the relevant literature, is a frequently observed situation that should be investigated in other DIF studies to be performed in the future. Besides, studies analysing the cognitive levels in speed-time problems and the solution strategies used by applicants of different gender groups could illuminate this point.

| | |
|---|---|
| If<br>$x + \frac{1}{x} = 3\sqrt{5}$ ,<br>then which of the following is equal to<br>$\left(x - \frac{1}{x}\right)^2$ ?<br>A) 37 B) 39 C) 40 D) 41 E) 43 | If a vehicle moving from city A to city B travels at 80 kmph, it arrives at city B 5 minutes later than it is supposed to, and if it travels at 100 kmph, it arrives 20 minutes earlier.<br>How many minutes is it supposed to take for this vehicle to go to city B?<br>A) 120 B) 60 C) 40 D) 30 E) 20 |

Figure 3. The Two Items Functioning in favour of Female Applicants

Expert opinion was consulted so as to investigate the causes for DIF displaying items to function differently according to gender groups. For this purpose, DIF displaying items and the ways they displayed DIF were presented to the experts and their views on the causes for different functioning were asked. The views and the items were analysed. In accordance with the experts' views, it was concluded that all the factors likely to be the causes for DIF had remained within the mathematical/ quantitative ability construct which was intended to be measured in the tests. On considering the DIF displaying items in Quantitative1 and Quantitative 2 tests according to gender, it was found that the items which were expressed abstractly in algebraic terms and which could be solved through algorithmic operations functioned in favour of female applicants while the items which were expressed as real-life problems and which could not be solved through routine algorithmic operations were in favour of male applicants. Some of the experts included in the study stated that female applicants might have been done better than male applicants at equal ability levels due to their tendency to carry out the operations regularly and step by step. Kalaycıoğlu and Kelecioğlu (2011) also reported similar findings. Accordingly, it was stated that male applicants perceived mathematics as a concept more valuable and more usable in their life than female applicants did (Fennema & Sherman, 1977). This situation might have caused male applicants to be better at practical problems taken from real life than female applicants at equal ability levels. Skills such as carrying out the operations step by step and regularly- which emerge as a factor functioning in favour of female applicants- and solving real life problems by means of mathematical models- which emerge in items functioning in favour of male applicants- can be considered as skills included in the construct which is intended to be facilitated with mathematics education and to be measured with tests.

### *Findings for DBF Analysis*

Bundles of items likely to display DBF were formed by considering the findings coming from DIF analyses and relevant literature. Because DBF analyses were conducted after DIF analyses, initial hypotheses about the groups the bundles would function in favour of were not established; instead, only findings obtained from DBF statistics were given. The six bundles formed are described below. The bundles formed on the basis of certain properties had intersection points. That is to say, some of the items belong in more than one group.

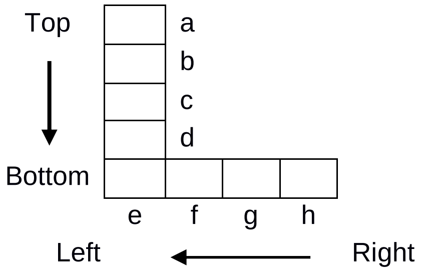
*Operations*. Items expressed abstractly only in algebraic and numerical terms were grouped as operational items. It was found in this study that the majority of the items functioning in favour of female applicants were of this type. Moreover, Bakan Kalaycıoğlu and Kelecioğlu (2011) also report DIF findings in favour of female applicants in some of such items. Similar findings were also reported by other researchers (Bakan Kalaycıoğlu & Berberoğlu, 2010; Cohen & Ibarra, 2005; Harris & Carlton, 1983).

*Word problems*. This bundle of items was composed of problems that presented real-life situations, in which the data were not presented in tables or charts but were presented verbally. Such items were found among the items displaying DIF in favour of male applicants. Studies are available indicating that word problems display DIF in favour of male applicants (Bakan Kalaycıoğlu & Berberoğlu, 2010; Harris &Carlton, 1983; Mendes-Barnett & Ercikan, 2006).

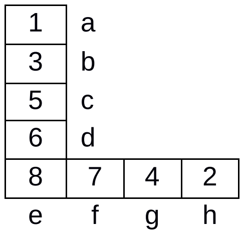
*Geometry*. Items requiring knowledge of geometry were included in this bundle. Contrasting research findings are available in geometry items displaying DIF according to gender (Berberoğlu, 1996; Cohen & Ibarra, 2005; Doolittle & Cleary, 1987; Mendes-Barnett & Ercikan, 2006). Geometry items did not display remarkable DIF according to gender in this study.

*Analytic reasoning*. It is the type of item used only in ALES among the examinations administered across Turkey. It has not been described as a subject domain in the primary or secondary school mathematics curriculum. Items of this type do not require knowledge of a special mathematical subject domain, but they can be answered by solving the given situation analytically. They can be likened to puzzles. Graduate Record Examinations (GRE), examinations similar to ALES in the USA, used to contain a sub-test of such items. Yet, GRE analytic reasoning skills test was no longer a multiple-choice test following the year 2002, and it was replaced by open-ended items measuring critical thinking and analytical writing skills (Educational Testing Service, 2007). Two examples are given for this bundle of items below in Figure 4.

Answer the following two items according to the information below.



The boxes in the above arrangement are named with the letters a, c, d, e, f, g, h. The numbers from 1 to 8 are used once and placed in the boxes, increasing both from top to bottom and from right to left. An example arrangement could be as follows.



As seen in this example, the numbers increase both from top to bottom and from right to left.

Given the number placed in box d is 4, what number is placed in box h?
A) 2 B) 3 C) 5 D) 6 E)7

Which box has the same number in all the possible arrangements?
A) a B) b C) c D) d E) e

Figure 4. Two Examples for Items of Analytical Reasoning

*Items which can be solved by trying numbers*. This bundle of items contains items in which answers can be found by trying the numbers given in options or the numbers probable to be appropriate. It was seen in DIF findings according to domains that the items functioning in favour of applicants of the verbal test had this property. Such items were also reported by Scheunemann and Grima (1997) to have functioned in favour of applicants of verbal tests. Items that could be solved by trying numbers were divided into two categories as operational items and problems.

*Items of secondary education (high school) curriculum*. This study found items requiring knowledge of subjects that were not included in the primary education mathematics curriculum but which were included in secondary education (high school) mathematics curriculum among the items displaying DIF in favour of applicants of the quantitative test. Such items were put under the heading of items of secondary education curriculum.

The item no and the number of items included in bundles of items and the results for DBF analyses conducted with SIBTEST are shown in Table 10. Statistical significance level was chosen as 0.01, and the groups to whose advantage the item bundles having $\beta_u$ statistics functioned are shown in the same table. Operational items displayed DBF in favour of female applicants and word problems displayed DBF in favour of male applicants in both tests. This was a finding in parallel to the ones obtained in DIF analyses and in the literature (Cohen & Ibarra, 2005; Harris & Carlton, 1983; Mendes-Barnett & Ercikan, 2006). It was found that geometry items in Quantitative 1 test did not display DBF but that they functioned in favour of male applicants in Quantitative 2 test.

Table 10. Item Number and Results for DBF Analysis for the Items in the Bundles

| Bundles of items | Item no | | Quantitative 1 | | Quantitative 2 | |
|---|---|---|---|---|---|---|
| | Quantitative 1 | Quantitative 2 | $\beta_u$ | Advantaged group | $\beta_u$ | Advantaged group |
| Operation | 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 16, 17 | 1, 2, 3, 4, 5, 7, 12, 13, 14, 15, 16 | -0.504 | Female | -0.422 | Female |
| Word problems | 15, 20, 21, 22, 24, 25, 26, 29, 30, 34, 35, 36, 37 | 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 32, 33, 34, 37, 38 | 0.446 | Male | 0.611 | Male |
| Geometry | 38, 39, 40 | 35, 36, 39, 40 | -0.032 | | 0.126 | Male |
| Analytic reasoning | 27, 28, 31, 32, 33 | 29, 30, 31 | 0.177 | Male | -0.031 | |
| Number trying | 2, 5, 9, 10, 13, 14, 24, 25, 26 | 1, 2, 6, 8, 11, 21, 22, 23 | -0.078 | Female | -0.026 | |
| Secondary education curriculum | 3, 8, 20, 39 | 4, 6, 8, 12, 13, 14, 19, 24, 25 | -0.147 | Female | -0.051 | |

It was apparent that this situation that did not appear in DIF analyses in which items were considered individually appeared in consequence of the combination of DIF effects at low levels in the items. Another situation in which small DIF effects combine and become remarkable was apparent in analytical reasoning items in Quantitative 1 test. Those items as a bundle also functioned in favour of male applicants. Items in which knowledge of subject areas that were not available in the primary education curriculum was effective were found to be in favour of female applicants. Female applicants with ability levels equal to male applicants in Quantitative 1 test were found to have answered more items requiring knowledge of subject areas, while male applicants were found to have been better at items of analytical reasoning which did not require knowledge of subject areas. Bundle of items that could be solved by trying numbers in Quantitative 1 test functioned in favour of female applicants, whereas DBF findings for this bundle were not found to be significant in Quantitative 2 test.

## DISCUSSION and CONCLUSION

DIF analyses on the basis of gender demonstrated that 11 items had displayed DIF. Four of the items functioned in favour of male applicants, whereas seven items functioned in favour of female applicants. One of the items functioning in favour of male applicants was a problem of ordering rational numbers, while three were word problems in which real-life situations were given. Six items functioning in favour of female applicants were operational items which were given in abstract contexts and which could be solved with algorithmic operations. One item functioning in favour of female applicants, on the other hand, was a problem of speed. Findings concerning DIF according to gender were in parallel to the ones reported in previous studies (Bakan Kalaycıoğlu & Berberoğlu, 2010; Harris & Carlton, 1983; Mendes-Barnett & Ercikan, 2006). It may be generally said on the basis of DIF analysis results that the applicants of different gender groups having an equal number of correct answers in ALES Quantitative tests answered different items and that they had different answering patterns.

The results of DBF analyses demonstrated that the items displaying remarkable DIF at item level functioned in favour of groups. Operational items functioned in favour of female applicants in Quantitative 1 and Quantitative 2 tests. Word problems, on the other hand, functioned in favour of male applicants in both tests. Items that could be solved by trying numbers and the items requiring knowledge on subject areas in the secondary education curriculum functioned in favour of female applicants as a group. Analytical reasoning items, however, functioned in favour of male applicants. The fact that the final three groups of items function differently at the group level, although they did not display remarkable DIF at item level individually was the result of small DIF effects in items coming together and thus becoming more remarkable.

DIF analyses should be performed on all large-scale examinations as a routine and especially the sources of differential item functioning should be detected. Thus, efforts should be made to attain unbiasedness- an important component of the validity of tests administered. Due to the fact that different types of items can display DIF in different examinations, those analyses should be conducted for every examination in itself and thus, efforts should be made for healthy generalisations. In addition to DIF analyses according to gender, DIF analyses according to the departments of graduation should also be performed in ALES, an examination for which university graduates of differing branches apply and which is used in selecting students for post-graduate education. DIF analyses according to departments of graduation for the tests- which are the subject matter of this study- can be found in the doctoral dissertation from which this study was produced.

The fact that items display DIF does not necessarily mean that those items should not be used in tests. Yet, a group of applicants can be in a more advantageous position than others if the number of items supporting them is abundant in a test. Therefore, the number of items providing different groups with advantages could be balanced. Studies revealing the extent to which the presence of DIF displaying items in tests influences individual score differences could be performed. Besides, the effects of the availability of DIF displaying items in tests on test validity could also be investigated.

Uncovering the strategies applicants of differing groups use in solving the items and analysing the differences could be useful in detecting the source of DIF. Applicants may be asked to think aloud and to solve the items in this way. The operations applicants use on test booklets in solving the test items can also bring their strategies into the light. In addition to that, their approaches towards different types of items and their calculations can also be requested and thus, analyses can be done. Additionally, technologies monitoring applicants' eye movements and recording them while they are solving the items can also be employed for this purpose. The differences between applicants' solution strategies- how they use the tables and charts in a test item, for instance- can be analysed and thus, the sources of DIF can be detected more clearly.

This is an exploratory study concerning ALES rather than a confirmatory study testing initial hypothesis constructed beforehand. The findings obtained in this study and the DBF hypotheses to be developed by other researchers on ALES could also be tested in a confirmatory approach. The findings obtained in several studies can be generalised more effectively in this way, thus the sources of DIF can be demonstrated more clearly and they can be offered to test developers.

$\Delta R^2$ –DIF statistics, which is used in DIF analyses along with the logistic regression method- is not adequate on its own in detecting DIF in items when cut-off points- which are commonly used in the literature- are used. Cut-off points that can be used in large-scale tests should be formed in $\Delta R^2$ statistics by considering the first type of error and statistical power balance. Effect size classification, which can be used in IRT-based DIF analyses, should be made.

## REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

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

Asil, M. (2010). *Uluslararası Öğrenci Değerlendirme Programı (PISA) 2006 öğrenci anketinin kültürler arası eşdeğerliğinin incelenmesi*. Yayınlanmamış doktora tezi, Hacettepe Üniversitesi, Ankara.

Bakan Kalaycıoğlu, D., & Berberoğlu, G. (2010). Differential item functioning Analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment, 20*(5), 1-12.

Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 36*(161), 3-13.

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-47.

Cohen, A., & Ibarra, R. A. (2005). Examining gender-related differential item funtioning using insights from psychometric and multicontext theory. In A. M. Gallagher ve J. C. Kaufman (eds.). *Gender differences in mathematics: An integrative psychological approach* ( pp. 143-171). Cambridge: NY.

Doğan, N., & Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 33*(148), 100-112.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*(2), 157-166.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOGMG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International, Inc.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*, 543-553.

Educational Testing Service. (2007). *The GRE® Analytical Writing Measure: An asset in admissions decisions*. Downloaded from www.ets.org/Media/Tests/GRE/pdf/gre_aw_an_asset.pdf

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal, 14*(1), 51-71.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*(1), 3-14.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurment, 40*(4), 281-306.

Gierl, M., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26–36.

Gök, B., Kelecioğlu, H. ve Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenzsel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 35*(156), 3-16.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berenzer, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249-266.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11), 182-188.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications: California.

Harris, A. M., & Carlton, S. T. (1983). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137-151.

Higaldo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

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

Jodoin, M. G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Downloaded from http://www.apa.org/science/programs/testing/fair-code.aspx

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy–Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293–312.

Ong, Y.M., Williams, J. S., & Lamprianou, I. (2011). Exploration of the validity of gender differences in mathematics assessment using differential bundle functioning. *International Journal of Testing, 11*(3), 271-293.

Oort, F. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6*(2), 150–166.

ÖSYM. (2008). *2008 Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES) Sonbahar Dönemi Kılavuzu*. www.osym.gov.tr adresinden indirilmiştir.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*(1), 73-90.

Scheunemann, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performace for female and Black examinees. *Applied Measurement in Education, 10*(4), 299-320.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology, 75*(5), 1350-1362.

Stout, W., & Roussos, L. (1995). *SIBTEST user manual*. Urbana: University of Illinois.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Thissen, D. (2001). *IRTLRDIF v.2.0.b: Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio tests for differential item functioning*. Downloaded from http://www.unc.edu/~dthissen/dl.html

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

Waller, N. G. (1998). EZDIF: Detection of uniform and non-uniform differential item functioning with the Mantel-Haenszel and logistic regression procedures. *Applied Psychological Measurement, 22*(4), 391.

Wang, W., & Yeh, L. Y. (2003). Effects of anchor item methods on differential ıtem functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479-498.

Yıldırım, H. H., & Berberoğlu, G. (1999). Judgemental and statistical DIF analyses of the PISA-2003 Mathematics Literacy items. *International Journal of Testing, 9*(2), 108-121.

Zieky, M. (1993) Practical questions in the use of DIF statistics in test development. In P.W. Holland ve H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): *Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

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

Zumbo, B. D., & Thomas, D. R. (1996, October). *A measure of DIF effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.