KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

**Received:** December 13, 2017
**Revision received:** July 16, 2018
**Accepted:** July 18, 2018
**OnlineFirst:** August 6, 2018


www.estp.com.tr
**DOI** 10.12738/estp.2018.2.0317 • April 2018 • 18(2) • 447–470

***Research Article***

# Using the Delphi Technique and Focus-Group Interviews to Determine Item Bias on the Mathematics Section of the Level Determination Exam for 2012*

Halime Yıldırım[1]
*Istanbul Medeniyet University*

Şener Büyüköztürk[2]
*Hasan Kalyoncu University*

**Abstract**

The aim of this study is to determine whether items from the mathematics section of the 2012 Level Determination Exam indicate item bias according to gender and school type. In particular, the process of item bias has been determined using the Delphi technique and focus group interviews. A two-stage mixed method research has been used for the study. While the first stage consists of identifying items that display differential item functioning (DIF) according to gender and school type, the second stage consists of determining the sources of DIF using the Delphi technique and examining through a focus-group interview which DIF sources lead to item bias. Mantel-Haenszel and logistic regression methods have been used for DIF analysis. While two items with significant DIF were detected according to gender, five items in favor of private schools were detected according to school type. In the process of item bias, the reasons why items display DIF have been determined using the Delphi technique, and 22 DIF sources were agreed upon. Finally, an expert panel was made to examine whether the DIF sources are grounds for item bias or not. According to the panel of experts, one item according to gender and two items according to the school type have been determined to show bias.

**Keywords**

Test bias • Differential item functioning • Delphi technique • Item bias expert panel

* This study is a summary of the graduate thesis titled " An Investigation of Item Bias of the Mathematics Subtest in 2012 year Level Determination Exam" prepared at Gazi University Educational Measurement and Evaluation Department.

1 **Correspondence to:** Halime Yıldırım, Educational Sciences, Educational Measurement and Evaluation Department, Faculty of Education, Istanbul Medeniyet University, Istanbul Turkey. Email: halime.yldrm@gmail.com

2 Educational Sciences Department, Faculty of Education, Hasan Kalyoncu University, Gaziantep Turkey. Email: senerbuyukozturk@gmail.com

**Citation:** Yıldırım, H., & Büyüköztürk, Ş. (2018). Using the Delphi technique and focus-group interviews to determine item bias on the mathematics section of the level determination exam for 2012. *Educational Sciences: Theory & Practice, 18,* 447–470. http://dx.doi.org/10.12738/estp.2018.2.0317

Obtaining as accurate an error-free measurement as possible is desirable for being able to obtain accurate information about the quantity of a characteristic being measured and for making proper decisions based on these measurement results. However, having errors in measurement results is inevitable in social sciences such as education and psychology (Tan, 2013). Having the true value of a feature observed in the measurement is desirable, but the actual value cannot be obtained directly due to various errors involved in measuring. Estimating the true value is attempted with the help of observed values. The true value of the measured feature is the average of the scores obtained from an infinite number of measurements of the feature according to classical test theory (CTT) (Crocker & Algina, 1986). According to CTT, the observed scores of an individual selected without unbiased from the population is the sum of the true score and the error score (Crocker & Algina, 1986, pp. 5–6). The error mentioned here is a random error and can also be defined as the difference between the observed value and the true value for the individual being measured. Systematic errors are defined as consistent differences between groups unrelated to measured constructs or competences. The greatest difference between these two errors is that random errors concern the individual and systematic errors concern groups (Osterlind & Everson, 2009). The systematic errors involved in measurements for measurement and evaluation research in the field of education are generally called biases (Yurdugül, 2003). Additionally, in standards for educational and psychological testing, the concept of bias on a test means systematically higher or lower scores for a participating group with unrelated factors (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2011). Bias is also defined as involving systematic error based on the grouping of test scores from individuals in different subgroups (Camilli & Shepard, 1994; Zumbo, 1999). What is important in bias is that one of the subgroups of error sources (such as men or women) has an unfair advantage (Crocker & Algina, 1986; Zieky, 1993).

Bias has been examined under two headings: internal and external. According to Osterlind (1983), external bias occurs when the test scores of two or more groups have different level correlations with variables outside the test. External bias focuses on the predictive validity of a test, not the test items under the heading of test bias. Internal bias is defined as item bias (Atılgan, 2014). Two of the most important threats to validity according to Clauser and Mazor (1998) are item and test biases. Test bias is defined as invalid or systematic errors of test measurements for particular group members (Zumbo, 1999). In other words, it can be defined as tests used to estimate an interested construct having systematic above- or below-average estimations depending on the group. If a test is more advantageous for one group with a specific skill level than another group with the same skill level, and if the test is impacted by undesired sources, then the test is biased. These groups are often differentiated by

ethnicity, gender, native language, socioeconomic status, or disability. The reason for decreased validity of test scores or increased test bias has been defined as an item bias that works unfairly, favoring one group in the sense that they respond more accurately to a test item than other groups in most applications (Zieky, 1993). Item bias is the differentiation of the likelihood of two groups of the same skill level answering correctly due to features of the test items or test conditions that are inappropriate to the purpose of the test (Zumbo, 1999). If biased items are identifiable, the test can be made unbiased by removing them from the test. Thus, the vast majority of bias studies focus on test-item bias (Atılgan, 2014).

Item-bias analysis first focuses on whether each test item behaves similarly for the different subgroups obtained from the same population. Two issues exist at the center of the discussion on item bias. The first is participants' performance on an item being affected by other undesired sources of change compared to actual differences in the related construct. The second is whether the sources of undesired change that affect performance lead to systematic differences in certain subgroups of participants (Osterlind & Everson, 2009). For this reason, bias studies are carried out in line with two objectives by taking these subjects as the focus. The first is to determine whether the various subgroups are systematically affected by different sources of variance in the test. Second, if test scores are influenced by the same sources of variance for all subgroups, the investigation turns to whether these unrelated sources provide an unfair advantage to certain subgroups.

Biased items lead to differences in the probability of correct responses to an item from individuals in subgroups with the same skill level but who differ in terms of variables such as gender or socioeconomic level (Camilli & Shepard, 1994; Zieky, 1993; Zumbo, 1999). In the first step of determining item bias, differential-item-functioning (DIF) analysis is performed. DIF analysis is a kind of evidence-gathering process for item bias and a necessary but insufficient condition for item bias. The existence of difference does not indicate bias because the existing difference may be a real difference of ability, which is defined as item effect. In other words, two reasons exist for the emergence of DIF: item bias and the actual difference between subgroups, or item effect (Camilli & Shepard, 1994). In the case of item effects, a real difference exists between subgroups according to the construct the item measures; as such, the probability of correctly answering an item varies by subgroup (Camilli & Shepard, 1994; Zumbo, 1999).

One of the centralized national examinations in Turkey is the Level Determination Exam (LDE). Ensuring that large-scale exams such as the LDE have specific objectives like selection and placement is very important, as the consequences can affect individuals' future situations. Because of the measured non-specific qualities

of test groups, the ability of test items to gain any advantage from subgroups is important in terms of the accuracy of test scores and the conclusions drawn as a result. For this reason, examining whether the items of such tests show DIF is necessary in accordance with various individual characteristics as well as for removing items with bias from a test (Kan, 2007). The process of investigating item bias involves both performing statistical analyses (DIF analyses) and determining whether the identified items show DIF sources, either from actual differences measured or from undesired sources of variance. Whether the item showing DIF is biased or not is decided after being reviewed in terms of construct and content (Zieky, 1993).

Pre-applying exams such as the LDE, which is applied to secondary-education institutions for student selection and placement, over another group and making corrections regarding items on the tests compared with the application is not possible. This study is believed able to provide important information for informing testers about DIF and its potential sources, for drawing attention to its importance, and for elaborating on what kinds of DIF sources will cause item bias. In this light, knowing the factors that cause item bias when in the process of developing items will contribute to making more accurate decisions for evaluating students and tests used for this.

Based on DIF analyses from CTT, the Mantel-Haenszel (Holland & Thayer, 1988) and logistic regression (Swaminathan & Rogers, 1990) techniques are quite powerful approaches that have been used in identifying DIF in dichotomous items (Holland & Thayer, 1988). This study investigates whether or not the 8th-grade LDE mathematics subtest from 2012 shows DIF using the Mantel-Haenszel and logistic regression methods based on the variables of gender and school type. Many test-bias studies in the literature using gender as a variable have determined boys' and girls' item performances to be systematically differentiated (Abedalaziz, 2010; Bakan Kalaycıoğlu, 2008; Berberoğlu, 1995; Çepni, 2011; Karakaya & Kutlu, 2012). Berberoğlu and Kalender (2005) stated big differences exist in terms of learning outcomes among school types, observed most especially in Turkey among all the OECD countries. Kelecioğlu, Karabay, and Karabay (2014) determined that 69% of DIF items on the LDE show DIF in favor of pupils enrolled in private schools. Some studies on the variables of school type and gender determined the influence of a source outside the test to cause the differences between groups; other studies stated the differences to source from the scope the test wants to measure (Bakan Kalaycıoğlu, 2008; Bekçi, 2007; Yurdugül, 2003). On the other hand, answering whether or not results that show student performance on such tests differ according to school type and gender to be due to actual performance differences is also on the agenda. For this reason, analyzing bias in items showing DIF according to gender and school type is important for the validity of the tests.

In determining bias, examining whether the test content has been critiqued from different perspectives; whether it is a representation of a particular group or educational, professional, and racial roles; and its statistical techniques is also necessary. The second step aims to determine whether the group difference for items indicating DIF indeed result from talent or from measurement. The source of the difference is researched to determine whether items displaying DIF are biased or not. One method often used to determine whether items showing DIF are biased is to receive feedback from field experts through questionnaires for such items (Ateşok Deveci, 2008; Bakan Kalaycıoğlu, 2008; Çepni, 2011; Karakaya & Kutlu, 2012). By analyzing the opinions received from experts through the questionnaire, comments on the bias of the items can be made. The opinions obtained by experts from a one-time instrument can be regarded as a limitation, especially given that there is no opportunity to reconsider the item's sources. In this study, the DIF sources of items displaying DIF were determined using the Delphi technique. The Delphi technique is a research technique used by experts to identify, learn, and share the ideas of experts by searching for agreement among experts. Estimates are made by making decisions based on the opinions of expert panelists participating in the study. This research technique is based on the opinions of expert panelists rather than an individual's limited view (John, 2011). The use of the Delphi technique in this study seems important in terms of taking opinions from experts in a certain order, which leads to differences of opinion on the source of the problem disappearing early, along with convergence of the experts' views. After identifying the DIF sources using the Delphi technique, focus-group interviews were used to examine whether the identified sources lead to item bias. During the focus-group interview, items displaying DIF were examined in terms of structure and content to identify whether the sources determined by the Delphi technique cause item bias. The focus-group interview technique has been preferred for this purpose as it enables deeper and richer information to be reached than from individual views. In other words, identifying DIF sources using the Delphi technique and whether the determined sources lead to item bias has been examined as a further dimension of the study's originality using the focus-group interview within a qualitative study. This is considered an option for methodologies in the field. The literature has no research on the use of the Delphi technique and focus-group interviews as different methods in studies on item bias. The use of different techniques for this process seems important in order to gain a more methodological approach to the task of determining item bias. In this regard, the study is a first. The fact that expert opinions will be taken in this study using the Delphi technique is not only important in terms of the reliability of the results but also in eliminating this limitation, not leading to a statistical agreement in experts' opinions.

The purpose of this study is to determine whether items on the 2012 LDE mathematics subtest show DIF according to the variables of gender and school type,

what the DIF sources of items displaying DIF are using the Delphi technique, and whether these sources lead to item bias using the focus-group interview.

## Method

### Research Design

The study was conducted in two stages that take into account the nature of item-bias studies (see Figure 1).

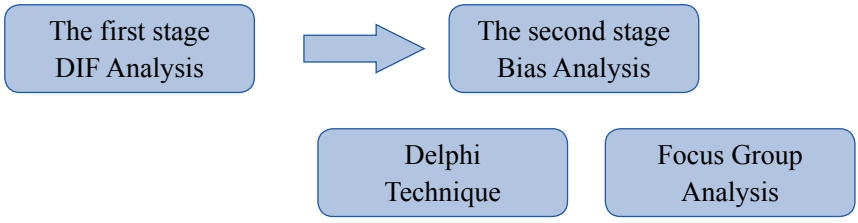


*Figure 1.* Research process.

DIF analysis is conducted in the first stage of the study for the 8th-grade LDE mathematics subtest from 2012 according to the variables of gender and school type. In the second stage, data are first obtained on identifying the reasons for DIF using the Delphi technique and then whether they cause item bias using the focus-group interview. The study can be described as a descriptive mixed-pattern as it first requires quantitative research, then qualitative research, and finally interprets these results together (Büyüköztürk, Çakmak, Akgün, Karaden, & Demirel, 2014).

### Population

The population of the research consists of the 1,075,546 students who had received at least one correct answer from the 2012 LDE mathematics subtest organized by the Ministry of National Education's (MoNE) General Directorate of Innovation and Education Technologies. Because the data are stored electronically, the total number of students obtained after excluding data that lacks school type or gender from the data set became 1,063,570. The study's analysis was carried out on the population in order to prevent errors caused by sampling choice. DIF analyses were done according to the variables of gender (boys and girls) and school type (private school and state school). Female students constitute 50.7% and male students 49.3% of the students taking the LDE. Students in private schools constitute 3.5% and those in state schools 96.5% of the students taking the LDE.

**Data Collection Instruments**

Based on the scope of the study, the data collection process consists of two stages: analysis regarding identification of items displaying DIF and their reasons why they display DIF, as well as examining them for item bias. In the first stage, student data from the eighth-grade LDE mathematics subtest from 2012 have been used for DIF analyses. The data used in the second phase have been obtained from experts regarding the bias status of items identified as having DIF. The second stage consists of two further sub-stages: the Delphi technique and the focus-group interview.

Circular No. 2008/77 from 11/12/2008 issued by MoNE stated aiming to measure only the subjects for that year and the curriculum achievements in the 6th-, 7th-, and 8th-grade LDEs. According to the e-application manual (2012) for the transition system to secondary-education institutions, the Level Determination Exam's mathematics subtest as applied on the LDE to 8th-grade students consists of 20 multiple-choice questions, and the 8th-grade mathematics items were prepared for measuring curriculum outcomes. The mathematics subtest's mean and standard deviation scores are 4.39 and 5.77, respectively.

**The Delphi technique and implementation: Identifying DIF sources.** The Delphi technique used to identify possible sources of DIF in the study consists of: identifying the problem, selecting the expert panelists, and preparing and applying the questionnaire forms used in the Delphi panels, as well as progressive and interactive phases. The Delphi technique is used as a means of consensus in cases where differences in opinion exist about similar situations and where decisions are likely to be influenced by strong groups (Helmer, 1967; Şahin, 2001). According to Linstone and Turoff (1975), Delphi is regarded as a technique that takes the group communication process as a whole and considers the group to be more effective than the individual in order to overcome a complicated problem. The process of structured communication takes place through the gathering of the group experts' opinions on a complex problem. Collecting opinions is aimed at establishing a common tendency using statistics and reaches consensus by applying sequential questionnaires (Şahin, 2001). While the Delphi technique is implemented with two, three, or in some cases more applications, the process continues until consensus is reached by passing the results of the application to each participant after each application (Dalkey, 1972; Şahin, 2000). The success of Delphi studies largely depends on the choice of the relevant experts in the field; selecting appropriate experts is seen as the most important feature enhancing the validity of Delphi studies (Delbecq, Van de Ven & Gustafson, 1975; Okoli & Pawlowski, 2004). In this research, experts who have important ideas on the research topic and are thought to contribute to the study have been chosen in terms of their experience and qualifications. In order to provide different views on the possible sources of DIF, three different specialist groups that include mathematics

teachers at MoNE, teachers working in private schools, and research staff participated in the study. Three Delphi panels were made respective of these groups, and three special questionnaires were applied to each panel.

The purpose of the first Delphi questionnaire is to obtain expert opinions on the possible sources of DIF and includes explanations about DIF and item bias as well as open-ended questions regarding items displaying DIF. In the questionnaire, the possible DIF sources of the items have been determined by asking, "What are the reasons for this item displaying DIF according to gender/school type?" The second Delphi panel occurred with the second Delphi questionnaire, which had been prepared by combining the opinions gathered from the first Delphi panel that resembled each other. The second Delphi questionnaire presented the experts with all the expert opinions regarding reasons for DIF being displayed; this was sent to experts in order to find out their level of agreement with these views and their order of importance regarding these views. Experts indicated their level of agreement for each item in the second questionnaire on a 4-point Likert-type scale (1 being "I definitely do not agree" and 4, "I definitely agree.") The purpose of the third Delphi survey is to inform the experts about the statistical results (general tendency of the group) from the second questionnaire and to have them reassess their level of agreement indicated in the second round. The content of the questionnaire is the same as from the second Delphi survey. The third Delphi questionnaire is tailored to each expert and additionally includes the agreement levels and statistics (percentage, median, and interquartile range) for each item as specified by the experts in the second round. At the end of the third Delphi panel, the reasons for showing DIF regarding strong and partial consensus were determined while items outside the criteria for compromise were excluded from the reasons for showing DIF.

**Focus group interview and implementation: item bias expert panel.** Whether the reason for items showing DIF as identified using the Delphi technique can be called item bias or item effect has been analyzed with a focus group interview designed as an expert panel.

All experts were preferred to be graduate students of measurement and evaluation who had graduated from mathematics teacher education departments, as the process of determining item biases requires mastery of both mathematics education and item-bias determination. The focus group interview session was realized with the participation of five experts and one director. Before the focus group interview session, information about the purpose of the session was given and the rules to be considered regarding the discussion process were explained. The focus group interview provided an opportunity for consensus for each of the DIF sources whose item had been identified as "DIF based on gender/school type," asking if the DIF source leads to

item bias. The session began and was directed by asking different questions in the direction of the given answers. The focus-group interview progressed with a focus on two issues related to items and DIF sources: The first is the aim/outcome that each item wants to measure while the second is whether the DIF source related to the item reflects a skill other than intended. Voice recordings were made during the session, and the recorded verbal data was transcribed. The session was completed in about 75 minutes upon accepting the justified views and decisions on bias as to whether a DIF source causes item bias.

## Data Analysis

The analysis process of the study consists of the steps for determining DIF and examining the biases of DIF-indicating items. Prior to DIF analyses, confirmatory factor analysis of unidimensionality was performed on asymptotic covariance matrices to test the unidimensional nature of the mathematics subtest data. The consistency of the model for the mathematics subtest was examined and determined that the mathematics subtest provides a sufficient level of model data adaptation. In other words, the test has been identified as unidimensional. In addition, the reliability coefficient (KR-20) calculated for the test is seen to be 0.85. The following steps have been followed in the analytical process applied at each stage.

**Differential-item-functioning analysis.** Whether the test items show DIF or not has been investigated using the Mantel-Haenszel (MH) and logistic regression methods in the context of CTT. The MH method is based on the chi-square statistic and can determine whether a relationship exists between the *MH* and $\chi^2$ statistics with item performance and group membership. However, the chi-square test does not give information about DIF size, which is the power of the relationship. The common likelihood ratio has been estimated for this. As interpreting the $\hat{a}_{MH}$ value that gives the size of the DIF is difficult, this is converted to a logarithmic scale $MH_{D\text{-}DIF}$ where interpretations are easier and more practical (Holland & Thayer, 1988). $MH_{D\text{-}DIF} = 0$ shows DIF to not be present; $MH_{D\text{-}DIF} > 0$ shows the item in favor of the focus group and $MH_{D\text{-}DIF} < 0$ indicates the item displays DIF in favor of the reference group (Zieky, 1993). Also, $MH_{D\text{-}DIF}$ statistics are used by the Educational Testing Service (ETS) as an effect-size statistic to interpret the size of the DIF (Wiberg, 2007). In this study, the program EZDIF (Waller, 1998) has been used for the MH analysis with the classification of *MH DIF* statistics as proposed by Zieky (1993). When used to determine DIF, items showing moderate (B) and high (C) DIF have been selected for analysis. The logistic regression method determines the strength of both uniform and non-uniform DIFs, the ability to test for significance, and the effect-size statistic (Clauser & Mazor, 1998). When performing DIF analyses using the logistic regression method, the variables are modeled hierarchically. DIF analyses are run by comparing the three models based on hypothesis testing. The models to be

compared have been called Model I, Model II, and Model III, respectively, according to the variables of the criterion (total score), group membership, and total score for group interaction. The presence of DIF has been determined by comparing Model III, which includes all variables, with Model I, which only includes the criterion variable. To determine DIF size, the value received from Models I ($R_1^2$) and Model III ($R_3^2$) have been used. The classification used by Bakan Kalaycıoğlu and Kelecioğlu (2011) and Çepni (2011) determining the amount of DIF was used in the logistic regression. In classification with logistic regression, average B- and high C-level items displaying DIF biasedness are identified as $0.10 < |\Delta R^2| < 0.20$ and $0.20 \leq |\Delta R^2|$, respectively. When determining DIF through logistic regression, syntax and the program SPSS (Zumbo, 1999) were used.

**Analyzing the bias states of items displaying DIF.** Analyzing the state of biases for items showing DIF has been completed in two stages. The first stage uses the Delphi technique to determine DIF sources, while in the second stage, the item bias that leads to the item bias of the DIF sources was determined through a focus-group interview identified as an expert panel.

**Delphi technique analysis.** The Delphi technique is known for data collection and systematic analysis (Franklin & Hart, 2007). Data were collected using three different questionnaires on the panels of the Delphi technique, and analyses of the collected data were conducted separately for each questionnaire. In the first Delphi questionnaire, content analysis was conducted to reveal important themes from the responses to the open-ended questions. The second and third Delphi questionnaires were conducted to determine the level of consensus. Detailed information on the statistics and analysis used in the Delphi questionnaires are given below.

***Analyzing the first Delphi questionnaire.*** In the first Delphi questionnaire, content analysis was conducted to extract important themes from the open-ended questions used to identify possible sources of DIF. The main purpose of content analysis is to reach the concepts and relations that can explain the collected data and to organize similar data by combining them under specific concepts and themes (Yıldırım & Şimşek, 2013). In the content analysis results, three experts evaluated the clarity of each DIF source and other DIF sources.

***Analyzing the second Delphi questionnaire.*** Statistics for participation levels were calculated for the data from the second Delphi questionnaire. Hasson, Keeney, and McKenna (2000) stated no universally-accepted consensus criteria to exist because the degree and level of consensus used in Delphi studies depend on the breadth of the participant group, the purpose of the research, and the resources. Determining the level of consensus means to have a certain percentage of responses given for items remain within specified ranges (Scheibe, Skutsch, & Schofer, 2002). Different criteria such

as percentage of participation, width between quarters, mean, median, and standard deviation are used to determine the biased items in the literature on the Delphi area (Scheibe et al., 2002; Şahin, 2001). In this study, participation percentage, median, and interquartile range values have been used for determining the biased items. The arithmetic average value is very sensitive to data at the end points of the distribution, thus the median is preferred because it is not influenced by data at the extreme points.

***Analyzing the third Delphi questionnaire.*** The third Delphi questionnaire to be used in the third Delphi panel has been tailored for each specialist and includes statistics on experts' levels of agreement on the questionnaires and the level of agreement of all panel members on each item. In this study, two sets of consensus groups have been defined: strong and partial. A group with a strong set of consensus has an agreement level of at least 80%, median of at least 3, and width between quarters of at most 1. If the three criteria are met, a strong consensus is reached regarding reasons for showing DIF. The partial compromise criterion group has at least an 80% agreement level, and partial agreement is reached regarding the reasons for showing DIF if one of the other criteria ($M \geq 3$ or $IQR \leq 1$) is met. Analyses of the third questionnaires have identified strong and partial consensus as sources of DIF.

**Analyses of the focus group interview.** In the focus group interview, expert opinions were obtained on whether a DIF source is biased for each of the DIF sources that reached consensus through the Delphi technique. Analysis of the focus group interview records was transcribed, followed by a content analysis of the transcriptions. In content analysis, a process was followed in which similar data were brought together and organized within the framework of certain concepts (Yıldırım & Simsek, 2013), taking into account the given justified views of whether or not the items are biased. According to the analysis results, the DIF sources provided by the consensus in the Delphi technique were determined to cause or not cause item bias.

## Results

This part of the study has determined the 8th-grade LDE mathematics subtest from 2012 to show DIF according to the variables of school type and gender. Whether an item identified as showing significant DIF causes DIF sources to lead to item bias has been determined.

### Do Items Show DIF in the MH and Logistic Regression Analyses According to Gender and School Type in Mathematics Subtest?

The results of the MH and logistic regression analyses of the items in the mathematics subtest according to gender and school type are given in Table 1.

Table 1
*DIF Analyses Results According to Gender and School Type*

| Group | MH | | | Logistic Regression | | |
|---|---|---|---|---|---|---|
| | DIF Item | Level | Advantaged Group | DIF Item | Level | DIF direction |
| Gender | 4 | B | Female | 4 | B | Uniform DIF |
| | 19 | B | Male | 19 | B | Uniform DIF |
| School Type | 5 | B | Private | - | - | - |
| | 9 | B | Private | - | - | - |
| | 10 | B | Private | - | - | - |
| | 16 | C | Private | - | - | - |
| | 17 | B | Private | - | - | - |

When examining Table 1, only Items 4 and 19 are found to show significant levels of DIF (B and C) according to gender. According to the MH method, Item 4 shows DIF in favor of female students and Item 19 in favor of male students. When examining the logistic regression analysis results, Items 4 and 19 were determined to show uniform DIF in terms of DIF direction. When analyzing the results of DIF analysis in the table, no items show significant DIF through logistic regression, while Items 5, 9, 10, 16, and 17 show considerable DIF through the MH method. Item 16 shows DIF at the C level in favor of private school students, while DIF at level B shows a significant level of DIF in favor of private schools.

**According to Delphi Panelists, What Are the Reasons for Items Showing DIF?**

Three consecutive Delphi panels were made to determine the sources of DIF for items showing DIF based on gender and school type. The findings from all Delphi panels have been interpreted under separate headings.

**Results from the first Delphi panel.** For each of the 7 items in the mathematics subtest that showed moderate or high levels of DIF as a result of the first Delphi panel, experts were asked their opinions about the DIF sources with "What are the reasons for this item showing DIF according to gender / school type?" The first Delphi panel analysis and a sample of DIF sources identified for each item are given in Table 2.

Table 2
*Samples of DIF Sources Determined by the First Delphi Questionnaire*

| DIF Items | DIF source samples |
|---|---|
| **Gender** | |
| Item 4 | Female students are better than males at seeing details and thinking in detail |
| Item 19 | The games played by male students have improved their calculation skills such as their four operation skills (marble, 52 play cards, etc.) |
| **School Type** | |
| Item 5 | Concepts such as architect and field, which are more familiar to students studying in private schools, are in the foundation of the item. |
| Item 9 | More practice in private schools on subjects about translation, reflection, and rotation; students encounter such question more often. |
| Item 10 | Students in private schools, which mainly teach in English, are more familiar with repeating as a term. |
| Item 16 | Teachers working in private schools have field mastery, experience, and skills in using and teaching a greater variety of teaching materials. |
| Item 17 | The use of the concepts of "three crossroads" and "road separation" in the context of the item. |

When examining Table 2, while the DIF source for the 4th item identified to show DIF according to gender reflects differences in cognitive skills, the DIF source for Item 19 reflects male students' experiential differences and related skills development. The DIF sources exemplified for Items 5, 10, and 17, which show DIF according to school type, seem to be related to the familiarity of the concepts at the root of the item (architect, repeating number, and three crossroads). The DIF source for Item 9 is seen to be related to private schools' extra practices in the subject area. Finally, the DIF source for the Item 16 is concerned with the experience and skills of teachers working in private schools on field dominance, efficiency, and material use.

**Results from the second Delphi panel.** As a result of the second Delphi panel, the statistics (participation percentage, median, and interquartile range) for experts' agreement levels on DIF sources were determined. Table 3 gives DIF sources with the highest consensus values for DIF items and statistics on these sources.

Table 3
*DIF Sources Provided the Highest Consensus According to Gender and School Type*

| DIF Items | DIF Sources | Consensus Criteria | | |
|---|---|---|---|---|
| | | % | *M* | *IQR* |
| **Gender** | | | | |
| 4 | Female students are more likely than male students to pass to the abstract period. | 93.8 | 3 | 1 |
| 19 | Male students are more curious than girls about sports and games where scores are calculated, such as football and basketball, and use them in daily life. | 100 | 4 | 1 |
| **School Type** | | | | |
| 5 | Private schools spend more time on comprehension activities | 100 | 3 | 1 |
| 9 | Private schools have visual materials (such as symmetry etc.), technological tools (projection, smart board, etc.) and dynamic geometry | 100 | 4 | 0 |
| 10 | Public schools' lack of such questions in the main source books and private schools excessive use of such questions in supporting sources | 75,1 | 3 | 1 |
| 16 | Private schools spend more time on project-based work where daily life problems are used than public schools do. | 93,8 | 3 | 0.75 |
| 17 | Private schools express the item as a real life problem and present the problems in relation to daily life | 93,8 | 3.5 | 1 |

Note: % = Agreement percentage, *M* = median; *IQR* = Interquartile range

When examining the statistics on the DIF sources with the highest consensus values for Items 19 and 4, which had been determined to show DIF in Table 3, agreement percentages were found above 90% (*M* = 3; *IQR* = 1). All DIF sources with the highest consensus values seem to provide strong consensus criteria. Furthermore, when examining the contents of the DIF sources given in Table 3 for Items 4 and 19,

they can be said to be generally concerned with the developmental characteristics of students and their familiarity with the forms and similar calculations given in the item content based on experiential difference and interests.

Table 3 gives DIF sources with the highest consensus values and statistics on these sources for Items 5, 9, 10, 16, and 17, which had been identified as showing DIF relative to school type. When examining the DIF sources with the highest consensus values for Items 5, 9, 16, and 17, which were determined to show DIF for private school students given in Table 3, agreement percentages were found to exceed 90% ($M > 3$; $IQR$ varies). When examining the two DIF source statistics for Item 10 in particular, the agreement percentage is seen to be less than 80% ($M = 3$; $IQR \leq 1$). DIF sources for Items 5, 9, 16, and 17, which have the highest consensus values, provide strong consensus criteria, while DIF sources with the highest consensus values for Item 10 satisfy neither the strong nor partial consensus criteria. When examining the contents of DIF sources with the highest consensus criterion values given for Items 5, 9, 10, 16, and 17 in Table 3, application differences made in schools in general are observed related to the effective use of visual materials, technological tools, and dynamic geometry programs; familiarity with similar content and type of questions; different teaching methods; and teachers' effective techniques (familiarity with tangram materials as a real life problem).

**Results from the third Delphi panel.** Analysis of expert opinions from the third Delphi questionnaire has identified DIF sources that provide partial and strong consensus. Consensus was achieved by applying the second and third Delphi panels in 22 of the 54 DIF sources obtained as the result of the first Delphi panel. 32 DIF sources that were not able to provide strong and partial consensus were eliminated.

Table 4
*Agreement Reached by DIF Sources According to Gender and Consensus Criteria*

| Items | DIF Sources | Consensus Criteria | | |
|---|---|---|---|---|
| | | % | *M* | *IQR* |
| *Item 4* | -Female students develop abstract thinking earlier than male students. | 100 | 3 | 1 |
| | - The household items and games female students play help with conic perceptions used in the field | 87.6 | 3 | 1 |
| | -Female students are better than males at seeing details and thinking in detail | 86.7 | 3 | 1 |
| *Item 19* | -The +3 points in the item are similar to calculating scores in soccer matches (*male students can find/predict how many matches the team has won when the total number of matches played by the teams is known*) | 93.3 | 4 | 0 |
| | The games that male students play improve their calculation skills, such as their four operation skills (marbles, 52-card deck, etc.) | 86.6 | 3 | 1 |
| | -More male students are into competitive programs | 80 | 3 | 0 |
| | - Male students are more curious than girls about sports and games that calculate scores such as football and basketball and use them in daily life | 100 | 4 | 1 |

The results from the third Delphi questionnaire showing DIF sources and consensus criteria for Items 4 and 19 identified to exhibit DIF according to gender are given in Table 4.

Data analysis from the third Delphi survey resulted in consensus; a strong consensus was reached for all opinions identified as DIF sources. Three different DIF sources were identified as a result of the consensus for Item 4. The DIF source with the highest consensus in Item 4 is the statistic on the level of agreement over "female students develop abstract thinking earlier than male students" (Agreement = 100%; $M = 3$; $IQR = 1$). Fourteen different DIF sources were identified in Item 19, which showed DIF in favor of male students. Item 19 show the DIF source with the highest consensus to be "similarity between +3 points in the item with calculating soccer match scores" and has agreement level statistics of Agreement = 93.3%, $M = 4$, and $IQR = 0$.

Table 5 gives the agreed DIF sources and consensus criteria for Items 5, 9, 10, 16, and 17 which were identified in the results of the third Delphi questionnaire as showing DIF according to school type.

When examining Table 5, a strong consensus has been reached over all DIF sources identified as demonstrating DIF according to school type. All identified DIF items are in favor of private school students, with four different DIF sources identified for Item 5. The DIF source that provides the highest level of consensus in Item 5 is "Private schools have more visual materials (e.g., symmetry mirror), technological tools (.g., projection, smart board) and dynamic geometry programs (e.g., GeoGebra) and use them more effectively," and has agreement statistics as: Agreement = 93.3%, $M = 4$, and $IQR = 0$.

Three DIF sources for Item 9 were identified as providing agreement, and the DIF sources with the highest agreement are as follows: "Private schools have more visual materials, technological tools, and dynamic geometry programs and use them more effectively" and "Students in private schools use tangram materials in their lessons (students in public schools do not know this material)." These two DIF sources have the highest agreement for Item 9 with the following statistics on level of agreement: Agreement = 100%, $M = 4$, and $IQR = 0$.

A single DIF source was identified in Item 10 for agreement ("Public schools' lack of such questions in the main source books and private schools' extreme use of such questions in supporting sources"). Statistics for this are Agreement = 86.6%, $M = 3$, and $IQR = 1$. Three different DIF sources were identified for Item 16. The other four DIF sources did not achieve strong or partial consensus. The DIF source that provides the highest level of consensus for this item is "The item is expressed as a real life problem and private schools present problems in relation to daily life." Statistics for this are: Agreement = 93.3%, $M = 4$, and $IQR = 0$. Finally, four different DIF sources

Table 5
*Agreement Reached by DIF Sources According to School Type and Consensus Criteria*

| Items | DIF Sources | Consensus Criteria | | |
|---|---|---|---|---|
| | | % | *M* | *IQR* |
| *Item 5* | - Private schools have more visual materials, technological tools, and dynamic geometry programs and use them more effectively. | 93.3 | 4 | 0 |
| | -Private schools spend more time on comprehension activities. | 100 | 3 | 1 |
| | -Private schools have more applications on similar subjects and their students are familiar with these types of questions. | 86.7 | 3 | 1 |
| | -The item is expressed as a real life problem and private schools use problems related to daily life. | 93.4 | 3 | 1 |
| *Item 9* | - Private schools have more visual materials, technological tools, and dynamic geometry programs and use them more effectively. | 100 | 4 | 0 |
| | - Private schools have students use tangram materials in their lessons. | 100 | 4 | 0 |
| | - Private schools provide more practice on translation, reflection, and rotation; students encounter these types of questions more. | 86.7 | 3 | 1 |
| *Item 10* | Public schools' lack of such questions in the main source books and private schools' extreme use of such questions in supporting sources. | 86.6 | 3 | 1 |
| *Item 16* | The item is expressed as a real life problem, and private schools present problems related to daily life. | 100 | 4 | 1 |
| | - Private schools have students spend hours reading books; their students have better reading habits than public school students. | 93.3 | 3 | 1 |
| | -Private schools spend more time on project-based work with daily life problems more than public schools. | 100 | 3 | 1 |
| *Item 17* | - Private school students are familiar with computer and mobile games, which makes it easier to perceive item content being constructed in a game style. | 93.3 | 3 | 1 |
| | - The item is expressed as a real life problem and private schools present problems in relation to daily life. | 100 | 3 | 1 |
| | - Private schools have more opportunities to improve imagination with visual material, computer, internet, and video support. | 93.3 | 4 | 1 |
| | - Private schools provide more practice with different types of questions by narrating them as fables or emotions in the context of in-class and extracurricular activities. | 93.4 | 3 | 1 |

were identified for Item 17. The DIF source providing the highest agreement is "The item is expressed as a real life problem, and private schools present problems related to daily life;" consensus statistics are: Agreement = 100%, $M = 3$, and $IQR = 1$.

**Expert opinions on DIF source assessments in terms of item bias.** Whether or not the reconciled DIF sources resulting from the Delphi technique applications are grounds for item bias has been examined by a panel of experts on item bias. Focus-group interviews were held in this context. Items biased according to gender and school type are given in Figure 2. DIF sources causing item bias are given in Table 6.

**17.** A rabbıt at a 3-way intersection chooses one at random. Each road splits into 2 narrow roads that the rabbit chooses from at random and continues. What is the probability that the rabbit will encounter a tortoise waiting on one of the narrow roads

**19.** In a contest program, every correct answer is given +3 points and every wrong one -2 points. Aysun, a participant in this competition, answered all 5 questions. According to Aysun's score of 10 at the end of the contest, how many questions did Aysun answer correctly?

**9.** Of the numbered tangram shapes below, which two are reciprocal reflections of each other?

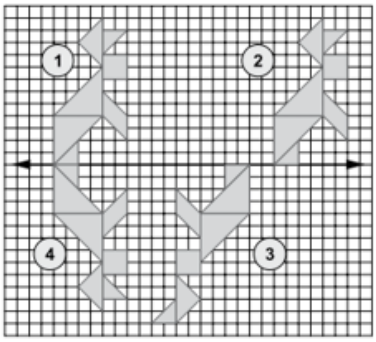


*Figure 2*. Biased items.

Table 6
*DIF Sources That Lead to Bias According to Gender and School Type*

| Items | DIF Sources | Advantaged Group |
|---|---|---|
| *Item 19* | The +3 points in the item resembles score calculations for soccer matches (male students can find / predict how many matches the team has won when the total number of matches the team has played is known) | Male |
| | -Male students are more into competition programs | Male |
| *Item 9* | - Private schools have students use tangram materials in their lessons (Public school students do not know this material) | Private school |
| *Item 17* | -The familiarity of private school students with computer and mobile games makes it easier to perceive the content of the item because it is constructed in the style of a game. | Private school |
| | -- Private schools have more practices for different types of questions narrated as fables or emotions in the context of in-class and extracurricular activities. | Private school |

When identifying all the DIF sources that reached agreement for Item 4 as item effect on DIF sources that had been determined for Items 4 and 19 according to gender, Table 6 shows two DIF sources mentioned for Item 19 that constitute bias. Experts who intend to measure DIF, which has been shown to demonstrate DIF in favor of male students, have indicated this item to require "solving linear systems of equations using algebraic methods." While four DIF sources had reached agreement for this item, experts in the focus-group interview saw two of these DIF sources to be caused by bias. The first DIF source stated the reason for bias as "the +3 points

resembles score calculations in soccer matches." Although the question needs to be solved using arithmetic for the equation system for the DIF source, giving the scope in the form of an account of +3 points and -2 points is considered to form a resemblance with soccer score calculations; more equal conditions could be created for female students. For this reason, the experts indicate the DIF source to be due to item bias. The DIF source specified in Table 6 as the reason for bias is "Male students are more into competition programs." In relation to the DIF source, stating the content of the question to be designed as a real life problem and this situation to be specially designed as a contest program, which is considered more exciting for male students, creates bias in favor of male students.

When determining as item effects all the agreed DIF sources for Items 5, 10, and 16 from among Items 5, 9, 10, 16, and 17 that had been identified as showing DIF according to school type, an examination of Table 6 shows bias for one DIF source from Item 9 and two DIF sources from Item 17. According to Table 6, the DIF source reported to lead to item bias for Item 9 is seen as "Students in private schools use tangram material in their lessons (students in public school do not know this material)." They indicated Item 9 to be a question that requires "determining a given translated reflection for a given figure." However, the tangram material at the root of Item 9 is seen as a reason for bias because private school students are more familiar with tangrams. They stated that "Using figures that both groups are familiar with, rather than tangram and the shapes that can be formed with this material" can prevent bias in this item. They expressed the opinion that the item's content "uses tangram materials at its base, and private school students are familiar with the material and with the question content." Experts also noted that "Using tangram materials that were used in classroom activities for learning subjects on transformational geometry subjects serves the purpose, but including this material as a fundamental of the question positively biases private school students."

Related to the reason for Item 17 showing DIF, experts indicated the item to require "solving questions about the possibility of an event." When examining Table 6, one source of DIF finding agreement for Item 17 is "private school students' familiarity with computer and mobile games makes it easier for them to perceive the item content because it is constructed in the form of game." Experts say that if the content of the item is effective in solving the problem, it causes item bias. In addition, they indicated that "if the item is structured in the form of a classical probability question rather than in a form of game, the item will have the same features for private and public school students." Similar comments have been made on the other DIF source commenting on the possibility of bias.

## Discussion

This study has investigated whether the 2012 LDE mathematics subtest shows bias according to the variables of gender and school type. The reasons for items displaying DIF show DIF at a particular level of significance have been identified using the Delphi technique, and item bias has been determined using a focus group interview designed as a panel of experts.

As a result of DIF analyses, the Items 4 and 19 show significant DIF by gender using the Mantel-Haenszel and logistic regression methods. Item 4, which favors girls, requires the creation of a surface expansion of a vertical circular cone, while Item 19, which favors boys, is modeled as a real life problem. When examining the DIF determination studies for gender using logistic regression and MH methods in the literature, findings similar to those obtained from this study were encountered (Çepni, 2011; Harris & Carlton, 1993; Li, Cohen & Ibarra, 2004; Yurdugül, 2003). The study of Berberoğlu (1995) determined that geometric items show DIF in favor of girls, which parallels the results of this study. Similarly, the fact that items given as real-life problems favors males is similar to findings obtained from studies done in Turkey and abroad (Bakan Kalaycıoğlu & Berberoğlu, 2010; Mendes-Barnet & Ercikan, 2006; Yurdugül & Aşkar, 2004). According to the results from the MH analysis according to school type, only Item 16 shows DIF at the C level; Items 5, 9, 10, 16, and 17 show significant levels of DIF at the B level in favor of private school students. Items 5, 16, and 17, identified as showing DMF in favor of private school students, are designed as real life problems. When examining the literature, many items identified as showing DIF, similar to this study, favor private school students (Bakan Kalaycıoğlu, 2008; Bekçi, 2007; Karakaya, 2012; Kelecioğlu et al., 2014; Özdemir, 2003).

As a result of the Delphi panels, a total of 22 DIF sources were identified for items as showing significantly higher DIF according to gender and school type. When examining the relevant literature, various studies have been conducted on DIF and its possible sources (Berberoğlu, 1995; Kurnaz, 2006; Stage, 1997; Yurdugül, 2003; Zenisky, Hambleton, & Robin, 2003). In particular, DIF sources in agreement on gender are generally grouped under the familiar titles of cognitive skills and developmental differences among gender groups, experiential differences, item content, and its related information. Similar to the findings obtained in this study, Bakan Kalaycıoğlu (2008) and Li, Cohen, and Ibarra (2004) identified male students as a DIF source to be better at subjects that require three-dimensional thinking skills. Kurnaz (2006) identified the word group *spider-man* in the context of item as a DIF source in his study, expressing that the group's having gone to the popular spider-man film at that time facilitated the perception of item content as a result of experiential differences. In addition, a large number of studies identifying DIF sources for gender showed in the results that item content and the concepts contained in item content

affect item performance between gender groups (Karakaya, 2012; Yurdugül, 2003; Zenisky et al., 2003). In addition, Karakaya and Kutlu (2012) stated in their study on item bias in a Turkish subtest that "raising fish in an aquarium" is a source of DIF in item content because male students are more familiar with the topic.

The agreed-upon DIF sources in the items identified as showing DIF according to school type have been gathered under the following headings: Frequent and efficient use of visual materials, Familiarity with technological equipment in item concepts and content, subject content being related to daily life and constructed as a real life problem, abundant applications made on related subjects and similar types of questions, and the difference in implementation duration for the review period. Item constructed as real-life problems have been stated in many studies investigating possible sources of DIF to affect the item performances of individuals with the same ability levels (Abedalaziz, 2010; Bakan Kalaycıoğlu & Berberoğlu, 2010; Berberoğlu, 1995; Çepni, 2011; Li et al., 2004). Bekçi (2007) determined in his study that private-school students' being more likely to apply similar types of questions in lessons is a DIF source in the mathematics subtest. In the same study, he also expressed that private-school students' having different opportunities affects item performance, which in turn serves as a DIF source. In addition, Karakaya and Kutlu (2012) stated in their study on item bias that private schools' emphasis on comprehension activities being greater than state schools' is a possible source of DMF.

The focus-group interview analysis showed that in Item 19, which showed DIF according to gender, the two DIF sources were the reason for bias; Item 9 had one DIF source causing item bias and Item 17 had two DIF sources as the reason for item bias. When examining studies of DIF analyses, a number of studies have been found to determine beyond the DIF sources whether these sources are item-bias or item-effect, (Ateşok Deveci, 2008; Bakan Kalaycıoğlu, 2008; Bekçi, 2007; Karakaya, 2012; Karakaya & Kutlu, 2012; Kelecioğlu et al., 2014; Yurdugül, 2003). When examining the current study's reasons for bias and related items, the use of tangram materials for Item 9 was determined to constitute bias for private school students. Private schools' greater emphasis on the use of such materials can influence students' success (Martini, 1995) as well as their familiarity. Similarly, Karakaya and Kutlu's (2012) study found male students to be more interested in everyday life than girls and that using concepts related to raising fish in an aquarium at the root of the item causes bias in that item. Additionally, Educational Testing Service (2002) described item content issues that need to be taken into account when writing items and testing and that should be especially minimized, such as military issues, sports information, unnecessarily difficult words, violence, and more. This study also has found that requiring the use of words that have been used within the article causes bias in visual texts to reflect students' familiarity levels. The reasons for the biases set forth in Item

17 have been determined to favor school type; namely, DIF sources were examined in the form of play or daily life, and these schools' use of excessive applications was incorrect. A similar finding in Berberoğlu's (2009) study on evaluating the LDE indicated an important question of validity in LDE to have emerged at the stage for designing item content in the context of daily life. When designing an item at this stage, only the items needed for the response were stated as needing to be brought to the foreground; all other given explanations and drawings would significantly affect the validity of the item's scope. Finally, Item 19, which favors male students, has been determined to have its bias arise from the way the material had been constructed. Findings that resemble this study have been examined. In Bakan Kalaycıoğlu's (2008) study, wherever the physical subtest used cars (which are closer to male students' interests) in the item content, these items were determined as a non-test factor unaligned with the purpose of the problem and the cause of bias.

## References

Abedalaziz, N. (2010). A Gender related differential item functioning of mathematics test items. *The International Journal of Educational and Psychological Assessment, 5*(1), 101–116.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2011). *Standards for educational and psychological testing* (6th ed.). Washington, DC: American Educational Research Association.

Ateşok Deveci, N. (2008). Üniversitelerarası Kurul yabancı dil sınavının madde yanlılığı bakımından incelenmesi [Examination of the Interuniversity Board foreign language test in the frame of item bias] (Doctoral dissertation, Ankara University, Ankara, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Atılgan, H. (Ed.). (2014). *Eğitimde ölçme ve değerlendirme* [Measurement and evaluation in education]. Ankara, Turkey: Anı Yayıncılık.

Bakan Kalaycıoğlu, D. (2008). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Item bias analysis of the University Entrance Examination] (Doctoral dissertation, Hacettepe University, Ankara, Turkey.) Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

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, 29*(5), 467–478. http://dx.doi.org/10.1177/0734282910391623

Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Item bias analysis of the university entrance examination. *Education and Science, 36*, 3–13.

Bekçi, B. (2007). *Orta öğretim kurumları öğrenci seçme ve yerleştirme sınavının değişen madde fonksiyonlarının cinsiyete ve okul türüne göre incelenmesi* [Examining differential item functions of the elemantary school student selection and placement examination according to gender and school type] (Masters's thesis, Hacettepe University, Ankara, Turkey.) Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Berberoğlu, G. (1995). Differential item functioning analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation*, *21*, 439–456.

Berberoğlu, G. (2009). Evaluation of Ministry of National Education Level Placement Exam (SBS) applications. *Cito Education: Theory and Practice*, *2*, 10–24.

Berberoğlu, G., & Kalender, I. (2005). Investigation of student achievement across years, school types and regions: The SSE and PISA analyses. *Educational Science and Practice, 4*(7), 21–35.

Büyüköztürk, Ş., Çakmak, E. Ç., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2014). *Bilimsel araştırma yöntemleri* [Scientific research methods]. Ankara, Turkey: Pegem Akademi Yayınları.

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

Çepni, Z. (2011). Değişen madde fonksiyonlarının SIBTEST, Mantel Haenszel, lojistik regresyon ve Madde Tepki Kuramı yöntemleriyle incelenmesi [Differential Item Functioning Analysis Using SIBTEST, MantelHaenszel, Logistic Regression and Item Response Theory Methods] (Doctoral dissertation, Hacettepe University, Ankara, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

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.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando: Holt, Rinehart and Winston.

Dalkey, N. C. (1972). *Studies in the quality of life: Delphi and decision making.* Lexington, MA: Lexington Books.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). Group techniques for program planning: A guide to nominal group and delphi processes. Glenview, IL: Green Briar. Retrieved from https://sites.google.com/a/umn.edu/avandeven/publications/research/group-techniques-for-program-

Educational Testing Service. (2009). *Guidelines for fairness review of assessments.* Princeton, NJ: Author.

Franklin, K. K., & Hart, J. K. (2007). Idea generation and exploration: Benefits and limitations of the Policy Delphi research metod. *Innovative Higher Education, 31*(4), 237–246. https://dx.doi.org/10.1007/s10755-006-9022-8

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. *Applied Measurement in Education, 6*(2), 137–151.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, *32*(4), 1008–1015.

Helmer, O. (1967). *Systematic use of expert opinions (Report No. P-3721).* California, CA: The RAND Corparation. Retrieved from http://www.rand.org/content/dam/rand/pubs/papers/2006/P3721.pdf

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

John, N. (2011). Using the Delphi technique in educational technology research. *TechTrends: Linking Research and Practice to Improve Learning*, *55*(5), 24–30.

Kan, A. (2007). Test Fairness: DIF Analysis Accros Gender and Department of H.U Foreign Language Proficiency Examination. *Eurasian Journal of Educational Research, 29*, 45–58.

Karakaya, İ. (2012). An Investigation of item bias in science and technology subtests and mathematic subtests in in Level Determination Exam. *Theory and Practice in Educational Sciences, 12*(1), 215–229.

Karakaya, İ., & Kutlu, Ö. (2012). An Investigation of item bias in Turkish subtests in Level Determination Exam. *Journal of Education and Science, 37*(165), 348–362.

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Investigation of Placement Test in terms of Item Biasness. *Elementary Online, 13*(3), 934–953.

Kurnaz, F. B. (2006). *Peabody resim kelime testinin madde yanlılığı açısından incelenmesi* [Assessing item biased of Peabody Picture Vocabulary Test] (Masters's thesis, Hacettepe University, Ankara, Turkey.) Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115–136.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications.* Boston, MA: Addison-Wesley.

Martini, M. (1995). Features of home environments associated with children's school success. *Early Child Development and Care, 111*(1), 49–68. http://dx.doi.org/10.1080/0300443951110105

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.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management*, *42*(1), 15–29. Retrieved from http://spectrum.library.concordia.ca/976864/1/OkoliPawlowski2004DelphiPostprint.pdf

Osterlind, S. J. (1983). *Test item bias*. Beverly Hills and London: Sage.

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

Özdemir, D. (2003). A study on the effect of two categorical and priori weighted scales on differential agent function in multiple choice tests. *Education and Science, 28*(129), 37–43.

Şahin, A. E. (2000). Competencies of the principals of elementary school. *Educational Administration in Theory and Practice, 6*(22), 243–260.

Şahin, A. E. (2001). Delphi technique and its use in educational researches. *Hacettepe University Journal of Education, 20*, 215–220.

Scheibe, M., Skutsch, M., & Schofer, J. (2002). Experiments in Delphi methodology. In H. A. Linston & M. Turoff (Eds), *The Delphi method, techniques and applications (*pp. 257–281). Retrieved from http://is.njit.edu/pubs/delphibook/

Stage, C. (1997, April). *Do males and females with identical test scores solve test items in the same way?* Paper presented at the 23rd Annual Conference of the International Association for Educational Assessment in Durban, South Africa. Em No 23. Umea: Umea university, Department of Educational Measurement.

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

Tan, Ş. (2013). Öğretimde ölçme ve değerlendirme KPSS el kitabı [Measurement and evaluation in teaching: KPSS handbook]. Ankara, Turkey: Pegem Akademi Yayınları.

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

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods.* Umea University (EM No. 60).

Yıldırım, A., & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri* [Qualitative research methods in social sciences]. Ankara, Turkey: Seçkin Yayıncılık.

Yurdugül, H. (2003). *Ortaöğretim Kurumları Seçme ve Yerleştirme Sınavı'nın madde yanlılığı açısından incelenmesi* [Examination of selection and placement examination of secondary education institutions in terms of item bias] (Doctoral dissertation, Hacettepe University, Ankara, Turkey.) Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Yurdugül, H., & Askar, P. (2004). Examination of secondary education institutions student selection and placement exam in terms of sex bias. *Journal of Educational Sciences and Practice, 3*(5), 3–20.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). DIF detection and interpretation in large-scale science asssessments: Informing item writing practices. *Educational Assessment*, *9*(1-2), 61–78.

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

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. Retrieved from http://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf