International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 765–780

https://doi.org/10.21449/ijate.1250358

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

# Purification procedures used for the detection of gender DIF: Item bias in a foreign language test

**Serap Buyukkidik**[1,*]

[1]Sinop University, Faculty of Education, Department of Educational Sciences, Sinop, Türkiye

**ARTICLE HISTORY**

*Received: Feb. 12, 2023*
*Revised: Oct. 21, 2023*
*Accepted: Nov. 03, 2023*

**Keywords:**
Mantel-Haenszel,
Simultaneous item bias test,
Lord's chi-square,
Raju's unsigned area,
Bias,
Item purification.

**Abstract:** In the current study, differential item functioning (DIF) detection using real data was conducted with the application of "Mantel-Haenszel (MH)", "Simultaneous item bias test (SIBTEST)", "Lord's chi-square", and "Raju's area" methods, both when item purification was carried out and when item purification was not. After detecting gender-related DIF, expert opinions were obtained for a bias study since it is important to conduct gender bias research in the English test. Additionally, in the relevant literature, there were some DIF studies, but not completely similar bias studies. The sample of the research consisted of 7,389 students who took the "Transition from Primary to Secondary Education Exam (TPSEE, referred to as "TEOG" in Turkish)" administered in April 2017. When gender-related DIF analysis was performed with the aforementioned four methods, the results were found to differ partially. DIF analysis results differed in different conditions based on whether item purification was performed or not. Furthermore, the detection of DIF was indicative of potential bias. In the second stage of the study, the opinions of seven experts were sought for item 11, for which DIF was detected at least at B level based on MH, SIBTEST. As a result of expert opinion, it was established that there was no bias based on gender in any of the items in the English test. It is advised that akin bias studies be carried out to enable test developers to be aware of characteristics that may result in item bias and construct unbiased items.

## 1. INTRODUCTION

There is much research on gender differences in foreign language testing. Gender differences in the acquisition of a second language are controversial and emerge as a prominent topic in the literature (Llach & Gallego, 2012). However, the main question to be asked in such research is “Are these differences due to the real differences in the measured trait of different gendered individuals?” or “Do these differences stem from item bias?”. These questions have rarely been asked by researchers who conduct gender differences in achievement research.

The fairness and validity of the test are threatened in a test consisting of biased items (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014). People differ in terms of many demographic variables such as culture, gender, language, and ethnicity. The educational and

*CONTACT: Serap BUYUKKIDIK ✉ sbuyukkidik@gmail.com Sinop University, Faculty of Education, Department of Educational Sciences, Sinop, Türkiye

**e-ISSN:** 2148-7456 /© **IJATE** 2023

psychological assessment of this wide variety of people must be carried out with the same precision and fairness across groups, regardless of their irrelevant characteristics (gender, etc.) (Sireci & Rios, 2013). Item bias or differential item functioning (DIF) affects test fairness (Khalid & Glas, 2014). DIF and bias are two separate concepts (AERA, APA, & NCME, 2014). DIF shows the difference in the probability of individuals at the same ability levels responding correctly to the item and differentiation as a function of group membership (Hambleton & Rogers, 1989; Holland & Wainer, 1993; Camilli & Shepard, 1994; Zumbo, 1999; AERA, APA, & NCME, 2014). There are two reasons for detecting DIF: item bias and item effect, which are the real differences between subgroups (Camilli & Shepard, 1994). In other words, the detection of DIF is not always an indicator of bias (AERA, APA, & NCME, 2014). Bias is the systematic error in the item and test performances of individuals in different subgroups depending on the subgroup they belong to (Osterlind, 1983; Crocker & Algina, 1986; Camilli & Shepard, 1994; Zumbo, 1999; AERA, APA, & NCME, 2014). In bias studies, DIF analyses are performed at the first stage, and then, important reasons for the item bias are found by the expert opinion method. While the detection of DIF is a statistical process, the detection of item bias is a conceptual process based on interpretation (Camilli & Shepard, 1994; Zumbo, 1999; Wiberg, 2007). Item bias studies date back to Alfred Binet's test of mental capacity in 1910 (Camilli & Shephard, 1994).

Although many studies detect DIF today, the number of bias studies is quite limited despite its long history. In bias studies, it is seen that DIF studies are conducted without item purification generally (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya, 2012; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019). However, no bias study has encountered real data that deal with how the results differ with and without item purification in the detection of DIF. In this respect, there arises a need to conduct such bias studies in the literature.

### 1.1. Differential Item Functioning and Item Purification

Although the first studies on DIF were conducted by Cardall and Coffman (1964) and Angoff and Ford (1973) in the 20th century (Holland & Thayer, 1986), fairness in educational and psychological measures in the 21st century is a current issue that researchers still give importance (Sireci & Rios, 2013). There are many methods for detecting DIF and estimating its size. Some of these methods include Mantel-Haenszel, SIBTEST, IRT methods, standardization, chi-square, Likelihood ratio test, Logistic Regression, b parameter indices, probability differences indices, IRT Likelihood Ratio Test (LRT), general IRT-LR, log-linear models, and Lord's chi-squared test (Wiberg, 2007). Since these DIF detection methods are based on different statistical bases, detecting DIF using different methods may lead to different results (Çepni & Kelecioğlu, 2021; Bakan Kalaycıoğlu, 2022). Regardless of which DIF detection method is used, there are two groups in DIF analyses, the “focal group and the reference group”, and different functionalization between these groups is considered.

DIF detection methods can be classified in terms of “parametric vs. non-parametric”, “matching variable: observed vs. latent”, “dichotomously vs. polytomously”, “measure and/or test of DIF”, “uniform vs. non-uniform DIF”, “handle the cut-off score or not”, “sample size” (Wiberg, 2007). In addition, DIF is divided into uniform and non-uniform. If the probability of answering an item correctly is in favor of a group at all skill levels, it is said to be uniform DIF. If the probability of answering an item correctly is in favor of a different group at different skill levels, it is said to be non-uniform DIF (Camilli & Shepard, 1994; Zumbo, 1999). For example, while uniform DIF is detected in the MH method, non-uniform DIF can be detected in the Crossing-SIBTEST (CSIBTEST) developed in addition to the SIBTEST (Wiberg, 2007).

Using different DIF detection methods can affect results. Another factor affecting the differentiation of results in the detection of DIF is item purification The indication that the element for which the DIF is not detected means that the DIF detected in that element causes a type 1 error. Item purification is an iterative process used to control the error rate and increase the power and precision of the results (Khalid & Glas, 2014). According to Lord (1980), eliminating DIF items in iterative and multiple stages purifies test scores and reduces power and Type 1 error. Fidalgo, et al. (2000) discovered that different purification types (three amounts of DIF "(10%, 15%, and 30% of DIF-items), three test lengths (20, 40, and 60 items)" under different simulation conditions (single-stage, two-stage, and iterative) investigated the effect of the MH DIF detection method on performance. Based upon the findings of their research, they stated that the two-stage purification process was more effective than the one-stage purification process. Wang and Su (2004) suggested that two-stage and iterative purification could be safely used to reduce the inflated Type 1 error as a result of the Monte Carlo simulation study. When the related studies were examined, some studies suggested the iterative purification process (Lord, 1980; Fidalgo et al., 2000; Wang & Su, 2004; Khalid & Glas, 2014), but some of the research showed that purification does not improve the detection of DIF (Magis & Facon, 2013), indicating that there is no definitive conclusion. In this respect, it is important to conduct studies that reveal how DIF detection is affected when purification is performed and when it is not. When the purification studies were examined, it was seen that there were mostly simulation studies for the MH method (Wang & Su, 2004; Fidalgo et al., 2000). Studies comparing DIF results with and without purification on real data were quite limited (e.g., Özdemir, 2015; Tunc et al., 2018; Soysal & Yılmaz Koğar, 2021).

There have been many studies on DIF detection in the literature. However, most of these studies compared at least two methods (e.g., Emily et al., 2021; Soysal & Yılmaz Koğar, 2021). Emily, et al. (2021) performed Monte Carlo simulation and examined Lord's Chi-square (LC), LRT, and MH DIF detection methods in terms of type 1 error and found that the MH method had the best performance in terms of type 1 error. Soysal and Yılmaz Koğar (2021), on the other hand, determined DIF based on Lord's $\chi^2$ and Raju's unsigned area methods. It was found that a limited number of bias studies are carried out in large-scale or national examinations (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya, 2012; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019). In the bias studies conducted the effect of item purification was considered. In this study, we aimed to conduct comparative DIF analyses based on MH, SIBTEST, Lord's $\chi^2$, and Raju's unsigned area methods in real data set when item purification was or was not performed, and to conduct bias studies by obtaining expert opinions on items containing at least moderate DIF in at least two methods. In this respect, the study differed from similar studies. It was thought that the research would contribute to the literature. The reason for using Lord's $\chi^2$ and Raju's unsigned area methods in this research is that although there have been DIF studies using these two DIF detection methods with item purification and without item purification (e.g., Özdemir, 2015; Tunc, et al., 2018; Soysal & Yılmaz Koğar, 2021), there are no bias studies in the literature that consider purification. In addition, Tunc, et al. (2018) found that these two methods were the most sensitive in the purification process, which was effective in the selection of these methods. In addition, MH and SIBTEST are among the most used DIF detection methods in the literature. That's why these two methods were included in the study.

This specific study aimed to investigate whether the "Transition from Primary to Secondary Education Exam (TPSEE, referred to as "TEOG" in Turkish)", which was administered in April 2017 showed a gender bias. For this purpose, DIF detection was first carried out and if no purification was carried out, a DIF detection was carried out. Then, expert opinion was sought for the bias study. The study attempted to answer the following questions within the scope of the research:

1) In the TPSEE 2017 English test without purification, which items were gender-related-DIF detected based on the MH, SIBTEST and Crossing-SIBTEST, Lord's $\chi^2$, and Raju's unsigned area methods?

2) When purification is performed, in which items of TPSEE 2017 English test, gender-related DIF detected based on MH, SIBTEST and Crossing-SIBTEST, Lord's $\chi^2$, and Raju's unsigned area methods?

3) Which items in the TPSEE 2017 English test are gender biased according to expert opinion?

## 2. METHOD

### 2.1. Participants

The sample of the research consisted of 7,389 8th-grade students who took the TPSEE administered in April 2017. While 3,606 of these students were females, the other 3,783 students were males. Table 1 shows descriptive statistics for the total group, the focal group, and the reference group. In this study, females were treated as the reference group and males as the focal group. While the mean of the reference group is 13.58, the mean of the focal group is 11.28. The average test score of the female students is higher than that of the male students.

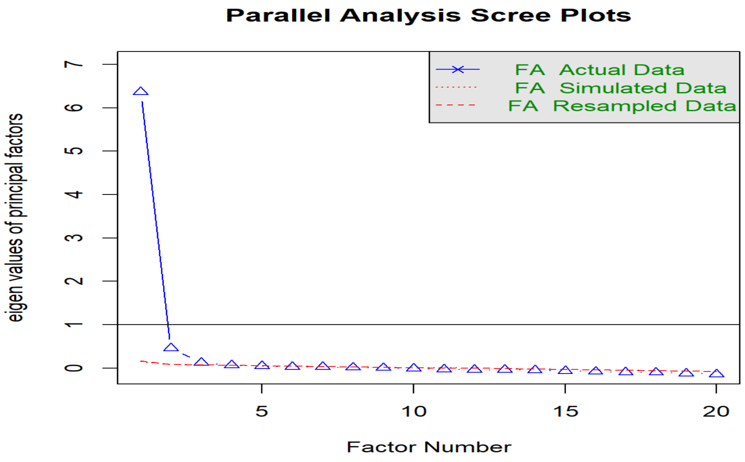
**Table 1.** *Descriptive statistics for reference and focal group.*

| | n | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|---|
| Total | 7389 | 0.00 | 20.00 | 12.40 | 12.00 | 5.55 |
| Reference group-Female (0) | 3606 | 0.00 | 20.00 | 13.58 | 14.00 | 5.35 |
| Focal group- Male (1) | 3783 | 0.00 | 20.00 | 11.28 | 10.00 | 5.51 |

### 2.2. Instrument

TPSEE is a national high school entrance exam for 8th-grade students, which was administered by the Turkish Ministry of National Education from 2013 to 2017. The exam consisted of six subtests including maths, science and technology, Turkish, English (as a foreign language), Turkish Revolutionary History and Kemalism, and religion, culture, and ethics. The data collection instrument was an English (foreign language) subtest (Booklet A) of TPSEE consisting of 20 multiple-choice items. The psychometric properties of the English subtest are discussed in this section. For the DIF analysis based on IRT, l, the unidimensionality assumption was first tested. As a result of the parallel analysis based on tetrachoric correlation, the English test was found to have a uni-dimensional structure (see Figure 1).

**Figure 1.** *Parallel analysis scree plots.*



When model data fit indices were examined, it was found that the data were suitable for one-dimensional structure ($\chi^2_{(170)}$ = 3760.91; $p$ = 0, RMSEA = 0.064, 90% CI [0.062, 0.065], TLI = 0.936). Factor 1 explains 49.95% of the variance. The factor loadings ranged from 0.47 (item 16) to 0.82 (item 18). The reliability of the measurements was good enough (KR-20 = 0.90, marginal reliability for 3 PL= 0.82).

When performing the item analysis in IRT, the model data fits in 1 PL, 2 PL, and 3 PL were examined. The model data fits for each model were treated according to five criteria. The model data fits for each model are presented in Table 2.

**Table 2.** *Model-Data Fit.*

| Model | AIC | AICc | BIC | SABIC | logLik |
|---|---|---|---|---|---|
| 1PL | 151026.19 | 151026.315 | 151171.253 | 151104.519 | -75492.095 |
| 2PL | 149191.191 | 149191.637 | 149467.501 | 149340.389 | -74555.595 |
| 3PL | 146169.156 | 146170.154 | 146583.62 | 146392.953 | -73024.578 |
| BEST | 3PL | 3PL | 3PL | 3PL | 3PL |

The 3 PL had the best model-data fit based on all five criteria (Akaike information criterion (AIC), Corrected AIC, Bayesian information criterion (BIC), sample-size adjusted BIC (SABIC), and likelihood ratio test (logLik)). Item parameters were obtained on the basis of 3 PL. Considering the discrimination parameter "*a*", it was seen that item 3 had the lowest discrimination ($a_3$=1.77), whereas item 17 had the highest discrimination ($a_{17}$=6.7). When the "*b*" item difficulty parameters were examined on the basis of 3 PL, item 13 had the lowest difficulty parameter ($b_{13}$=-1.25), while item 5 had the highest *b* parameter ($b_5$=0.68). When the "*c*" values or guessing parameters were examined, the item with the highest probability of answering by luck was item 16 ($c_{16}$=0.41), while item 3 had the lowest c parameter ($c_3$=0.01).

### 2.3. Analysis of Data

First, the missing data (111 student responses) were removed from the data set. The analysis of the data then began. In the analysis of the data, the construct validity and reliability proofs were collected in the first stage. In the second stage, unidimensionality and local independence from IRT assumptions were tested. Parallel analysis based on tetrachoric correlation was performed for unidimensionality (see Figure 1). For local independence, Yen's $Q_3$ index was examined and binary values were found below 0.20 in this research. To calculate the item parameters, analyses were performed according to 1 PL, 2 PL, and 3 PL. As the best model fit was achieved in the 3 PL, item parameters were considered according to the 3 PL. "irtGUI" package in R programming language (2021) was used for IRT assumptions (Yen's $Q_3$ index and parallel analysis) and for estimating marginal reliability.

In this study, DIF analyses based on 3PL were performed in IRT-based methods. After providing the assumptions, DIF analyses were performed with MH, SIBTEST, Lord's chi-square, and IRT Raju's unsigned area test. DIF analyses and parameter estimation based on IRT were performed using ShinyItemAnalysis (Martinkova & Drabinova, 2018) based on "mirt" (for item parameter estimation), "difR", and "ltm" (for DIF analysis) packages in R programming language (2021).

The code was generated for each method to indicate whether each method (MH, SIBTEST and Crossing-SIBTEST, Lord's $\chi^2$, and Raju's unsigned area) underwent purification separately, and the items found when purification was performed (1) and not performed (0) for 20 items. The agreement coefficient was calculated using ReCal (Freelon, 2013) to assess the agreement between the results.

In the final stage, a bias study was conducted by taking expert opinions on an item containing moderate DIF. Some characteristics of the experts whose opinions were used for the bias study are shown in Table 3.

**Table 3.** *Information about experts.*

| Id | Gender | Experience | Field | Id | Gender | Experience | Field |
|---|---|---|---|---|---|---|---|
| E1 | Male | 9 years | Measurement and Evaluation (MS, Ph.D.), Language Teaching (Undergraduate) | E5 | Female | 9 years | Measurement and Evaluation (MS, Ph.D.), Language Teaching (Undergraduate) |
| E2 | Female | 21 years | English Literature (Ph.D.) English Language Teaching (Undergraduate, MA) | E6 | Female | 11 years | English Language Teaching (Undergraduate, MS, Ph.D.) |
| E3 | Female | 15 years | English Literature (MA, Ph.D.) English Language Teaching (Undergraduate) | E7 | Male | 13 years | English Language Teaching (Undergraduate, MS, Ph.D.) |
| E4 | Male | 15 years | English Language Teaching (Undergraduate, MS, Ph.D.) | | | | |

Table 3 shows that a total of seven experts (two measurement and evaluation specialists and five English language teaching specialists) participated in the research. Four of these experts were lecturers in higher education who had completed their Ph.D. in English Language Teaching. While one of the other three experts was completing a Ph.D. in English Literature, two of them graduated from English language teaching and had completed their master's degree in measurement and evaluation and they were currently continuing their Ph.D. education. All experts had a master's degree in their field and had at least 9 years of experience. Descriptive data analysis was used for the qualitative part of the research in the bias study. The DIF detection methods used in the study are as follows.

### 2.3.1. *The Mantel-Haenszel test*

The Mantel and Haenszel test was developed by healthcare workers Mantel and Haenszel (1959), and its use in the detection of DIF is based on the work of Holland and Thayer (1986). The MH method uses the 2x2xK contingency table (Holland & Thayer, 1986). MH is a non-parametric, uniform DIF detection method based on classical test theory (CTT) that can be tested on polytomous and binary data (Wiberg, 2007). Zieky (1993) established reference ranges for the determination of DIF levels taking IΔMHI into account. When $|\Delta MH|<1$ there is no DIF or A-level ie negligible level. When it is in the range of $1\leq|\Delta MH|<1.5$, moderate (level B) DIF is detected. In the range of $|\Delta MH|\geq1.5$, a high-level (C level) DIF is detected.

### 2.3.2. *Simultaneous item bias test*

The SIBTEST for the detection of uniform DIF was developed by Shealy and Stout (1993) based on the standardization method. The basis of the CSIBTEST method developed to detect non-uniform DIF is based on the work of Li and Stout (1996), followed by Chalmers (2018). β values are obtained in the SIBTEST. When β is negative, the focal group is advantaged, and when β is positive, the reference group is advantaged (Gierl & Khaliq, 2001). In addition, Roussos and Stout (1996) suggested DIF classification according to the magnitude of the β value. When β is | β |<0.059, there is no DIF or A-level ie negligible level. When β is in the range of 0.059≤| β |<0.088, moderate (B-level) DIF is detected. If β is in the range of | β |≥0.088, a high level (C level) DIF is detected. SIBTEST is a method that allows the analysis of non-parametric binary and polytomous data (Wiberg, 2007).

### 2.3.3. *Lord's chi-square test*

Lord's chi-squared test is an IRT-based parametric DIF detection method based on Lord's work in 1980. It is an extended version of Lord's chi-squared test, the test of the b difference method, including the distinctiveness parameter (Lord, 1980). The Lord's chi-squared test allows the detection of DIF by taking into account the item parameters and the difference between the groups. Lord's chi-squared test is a method for detecting both uniform and non-uniform DIF that can be used in binary data (Wiberg, 2007). However, this method does not measure the size of DIF; it only tests for the presence of a DIF item (Wiberg, 2007).

### 2.3.4. *Raju's area method*

Raju's area method is based on Raju's work in 1988 and 1990. This method is a parametric method based on IRT. The logic of this method is based on the area between the item characteristic curves of the focal and reference groups in the signed area (Raju, 1988). In the null hypothesis, this area is equal to zero. Z statistics are used for this purpose. In recent studies, the unsigned area method is used. Raju's unsigned area is used to detect non-uniform DIF. Raju's unsigned area method is computed from the difference between the difficulty and discrimination parameters (Raju, 1988). One of the major problems with Raju's area methods (both signed and unsigned area) is their limitation for 3PL estimation. Raju (1988, 1990) showed that the area between two item response functions is infinite when the lower asymptotes are not equal. Raju (1988, 1990) suggested that equal or fixed c-parameters should be used for this situation. The guessing parameter *c* is estimated from the entire dataset and is considered fixed in the present study under the 3PL model.

## 3. RESULTS

In this section, the results of MH, SIBTEST, Lord's chi-square, and Raju's unsigned area method are given in terms of whether or not item purification was performed. DIF results are given for the 20-item English test.

### 3.1. DIF Results When No Item Purification was Performed

Table 4 shows the results of DIF analysis without item purification based on the MH method.

**Table 4.** *DIF results based on MH.*

| Item | *MH*($\chi^2$) | *p*-value | αMH | ΔMH | DIF Level | Advantage group |
|---|---|---|---|---|---|---|
| item1 | 7.26 | 0.01 | 0.84 | 0.40 | A | Male |
| item2 | 4.55 | 0.03 | 1.14 | -0.30 | A | Female |
| item3 | 6.89 | 0.01 | 1.20 | -0.42 | A | Female |
| item4 | 7.14 | 0.01 | 0.84 | 0.42 | A | Male |
| item5 | 17.44 | 0.00 | 0.77 | 0.62 | A | Male |
| item6 | 8.51 | 0.00 | 1.21 | -0.45 | A | Female |
| item7 | 11.80 | 0.00 | 1.28 | -0.59 | A | Female |
| item8 | 0.28 | 0.60 | 0.96 | 0.08 | - | - |
| item9 | 2.79 | 0.10 | 1.11 | -0.24 | - | - |
| item10 | 0.28 | 0.60 | 1.04 | -0.09 | - | - |
| **item11** | **55.69** | **0.00** | **1.73** | **-1.28** | **B** | **Female** |

| | | | | | | |
|---|---|---|---|---|---|---|
| item12 | 2.64 | 0.10 | 0.90 | 0.26 | - | - |
| item13 | 4.85 | 0.03 | 0.82 | 0.47 | A | Male |
| item14 | 0.14 | 0.71 | 1.02 | -0.05 | - | - |
| item15 | 0.03 | 0.87 | 1.01 | -0.03 | - | - |
| item16 | 6.21 | 0.01 | 1.15 | -0.32 | A | Female |
| item17 | 9.02 | 0.00 | 0.84 | 0.41 | A | Male |
| item18 | 15.92 | 0.00 | 0.76 | 0.64 | A | Male |
| item 19 | 0.09 | 0.76 | 1.02 | -0.05 | - | - |
| item 20 | 8.26 | 0.00 | 0.83 | 0.42 | A | Male |

When item purification was not performed in the MH method, DIF was detected in item 1, item 2, item 3, item 4, item 5, item 6, item 7, item 11, item 13, item 16, item 17, item 18, and item 20. Item 1, item 4, item 5, item 13, item 17, item 18, and item 20 contained negligible DIF in favor of male students. Item 2, item 3, item 6, item 7, item 11, and item 16 contained DIF in favor of females. Only a moderate DIF level was detected in item 11, e.i., a B level of DIF. A-level DIF was detected in all other items containing DIF.

Table 5 shows the results of DIF analysis without item purification based on SIBTEST and CSIBTEST methods.

**Table 5.** *DIF results based on SIBTEST and Crossing-SIBTEST.*

| Item | $\beta_{uni}$ | $\beta_{cro}$ | $X_{uni}^2$ | $\chi_{cro}^2$ | $p$-value$_{uni}$ | $p$-value$_{cro}$ | Uniform/Non-uniform | DIF Level | Advantage group |
|---|---|---|---|---|---|---|---|---|---|
| item1 | -0.03 | | 5.34 | | 0.02 | | Uniform | A | Male |
| item2 | | 0.03 | | 8.36 | | 0.02 | Non-uniform | A | - |
| item3 | 0.04 | | 11.68 | | 0.00 | | Uniform | A | Female |
| item4 | -0.03 | | 6.28 | | 0.01 | | Uniform | A | Male |
| item5 | -0.05 | | 16.98 | | 0.00 | | Uniform | A | Male |
| item6 | 0.03 | | 8.01 | | 0.00 | | Uniform | A | Female |
| item7 | 0.04 | | 15.46 | | 0.00 | | Uniform | A | Female |
| item8 | -0.00 | 0.00 | 0.08 | 0.08 | 0.77 | 0.77 | NO DIF | - | - |
| item9 | 0.03 | | 5.22 | | 0.02 | | Uniform | A | Female |
| item10 | 0.01 | 0.02 | 0.93 | 2.52 | 0.33 | 0.28 | NO DIF | - | - |
| **item11** | **0.08** | | **60.55** | | **0.00** | | **Uniform** | **B** | **Female** |
| item12 | -0.01 | 0.01 | 1.70 | 1.70 | 0.19 | 0.19 | NO DIF | - | - |
| item13 | -0.02 | 0.02 | 3.82 | 3.82 | 0.05 | 0.05 | NO DIF | - | - |
| item14 | 0.01 | 0.03 | 0.62 | 5.47 | 0.43 | 0.06 | NO DIF | - | - |
| item15 | 0.01 | 0.01 | 0.27 | 1.02 | 0.61 | 0.60 | NO DIF | - | - |
| item16 | 0.03 | | 6.95 | | 0.01 | | Uniform | A | Female |
| item17 | | 0.03 | | 9.85 | | 0.01 | Non-uniform | A | - |
| item18 | -0.04 | | 18.75 | | 0.00 | | Uniform | A | Male |
| item19 | 0.00 | 0.01 | 0.02 | 1.06 | 0.89 | 0.59 | NO DIF | - | - |
| item20 | | 0.03 | | 11.72 | | 0.00 | Non-uniform | A | - |

Table 5 shows that uniform DIF was detected in item 1, item 3, item 4, item 5, item 6, item 7, item 9, item 11, item 16, and item 18. Non-uniform DIF was detected in item 2, item 17, and item 20. Only item 11 contains DIF at the B level in the English test, while DIF at the A level was detected in the other items containing DIF. Item 1, item 4, item 5, and item 18 contained

DIF in favor of males. DIF was detected in favor of females in item 3, item 6, item 7, item 9, item 11, and item 16.

Table 6 shows the findings of DIF analysis without item purification according to Lord's $\chi^2$ and Raju's unsigned area methods.

**Table 6.** *DIF results based on Lord's $\chi^2$ ve Raju's unsigned area.*

| Item | Lord's $\chi^2$ | *p*-value | Raju's Z | *p*-value | Item | Lord's $\chi^2$ | *p*-value | Raju's Z | *p*-value |
|---|---|---|---|---|---|---|---|---|---|
| item1 | 3.52 | 0.17 | -1.87 | 0.06 | **item11** | **28.37** | **0.00** | **-4.77** | **0.00** |
| item2 | 22.88 | 0.00 | -4.44 | 0.00 | item12 | 2.36 | 0.31 | 1.56 | 0.12 |
| item3 | 0.74 | 0.69 | 0.81 | 0.42 | item13 | 20.50 | 0.00 | 2.86 | 0.00 |
| item4 | 7.97 | 0.02 | -2.86 | 0.00 | item14 | 0.04 | 0.98 | -0.19 | 0.85 |
| item5 | 2.82 | 0.24 | -1.68 | 0.09 | item15 | 0.58 | 0.75 | 0.67 | 0.51 |
| item6 | 9.27 | 0.01 | -2.81 | 0.00 | item16 | 11.31 | 0.00 | 3.41 | 0.00 |
| item7 | 3.24 | 0.20 | 1.78 | 0.07 | item17 | 13.23 | 0.00 | -2.96 | 0.00 |
| item8 | 0.01 | 1.00 | -0.09 | 0.93 | item18 | 12.86 | 0.00 | -3.71 | 0.00 |
| item9 | 6.42 | 0.04 | 2.53 | 0.01 | item19 | 4.65 | 0.10 | -2.16 | 0.03 |
| item10 | 3.41 | 0.18 | 1.78 | 0.08 | item20 | 14.70 | 0.00 | 3.88 | 0.00 |

Table 6 shows that DIF based on both Lord's $\chi^2$ and Raju's unsigned area methods was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, item 17, item 18, and item 20. DIF was detected in item 19 based on only Raju's unsigned area method.

It was seen that DIF was detected in item 2, item 4, item 6, item 11, item 16, item 17, item 18, and item 20 based on four methods when item purification was not performed. Only item 11 contained a moderate DIF in favor of girls. Other DIF-containing items contained negligible levels of DIF.

### 3.2. DIF Results When Item Purification was Performed

Table 7 shows the results of the DIF analysis based on the MH method when item purification was performed. When item purification was performed based on the MH method, it was detected that DIF was detected in item 1, item 3, item 4, item 5, item 7, item 11, item 12, item 13, item 17, item 18, and item 20. Item 1, item 4, item 5, item 12, item 13, item 17, item 18, and item 20 contained DIF in favor of males. Item 3, item 7, and item 11 contained DIF in favor of females. Only item 11 contained moderate DIF, while the rest of the DIF-detected items contained negligible (A-level) DIF.

**Table 7.** *DIF analysis results based on MH when item purification was performed.*

| Item | *MH*($\chi^2$) | *p*-value | αMH | ΔMH | DIF Level | Advantage group |
|---|---|---|---|---|---|---|
| item1 | 12.33 | 0.00 | 0.79 | 0.54 | A | Male |
| item2 | 1.12 | 0.29 | 1.07 | -0.16 | - | - |
| item3 | 4.26 | 0.04 | 1.16 | -0.34 | A | Female |
| item4 | 13.05 | 0.00 | 0.78 | 0.59 | A | Male |
| item5 | 35.54 | 0.00 | 0.68 | 0.90 | A | Male |
| item6 | 2.05 | 0.15 | 1.10 | -0.23 | - | - |
| item7 | 7.41 | 0.01 | 1.23 | -0.48 | A | Female |
| item8 | 3.39 | 0.07 | 0.88 | 0.29 | - | - |
| item9 | 0.00 | 1.00 | 1.00 | -0.00 | - | - |
| item10 | 0.00 | 1.00 | 1.00 | -0.01 | - | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| **item11** | **40.09** | **0.00** | **1.61** | **-1.12** | **B** | **Female** |
| item12 | 4.73 | 0.03 | 0.86 | 0.36 | A | Male |
| item13 | 4.33 | 0.04 | 0.83 | 0.45 | A | Male |
| item14 | 0.75 | 0.39 | 0.95 | 0.13 | - | - |
| item15 | 0.01 | 0.93 | 0.99 | 0.02 | - | - |
| item16 | 1.42 | 0.23 | 1.07 | -0.16 | - | - |
| item17 | 21.76 | 0.00 | 0.76 | 0.65 | A | Male |
| item18 | 27.42 | 0.00 | 0.69 | 0.86 | A | Male |
| item19 | 1.28 | 0.26 | 0.93 | 0.18 | - | - |
| item20 | 12.55 | 0.00 | 0.80 | 0.54 | A | Male |

In Table 8, DIF analysis findings in SIBTEST and CSIBTEST are given when item purification was performed. When item purification was performed, Table 8 shows that uniform DIF was detected in item 2, item 3, item 5, item 6, item 7, item 9, item 10, item 11, item 15, item 16, and item 18. Only item 11 contained DIF at the C level, while other DIF-detected items contained DIF at the A level. Item 5 and item 18 contained DIF in favor of males. DIF was detected in favor of females in item 2, item 3, item 6, item 7, item 9, item 10, item 11, item 15, and item 16.

**Table 8.** *DIF analysis results based on SIBTEST and Crossing-SIBTEST when item purification was performed.*

| Item | $\beta_{uni}$ | $\beta_{cro}$ | $X_{uni}^2$ | $\chi_{cro}^2$ | $p$-value$_{uni}$ | $p$-value$_{cro}$ | Uniform/ Non-uniform | DIF Level | Advantage group |
|---|---|---|---|---|---|---|---|---|---|
| item1 | -0.01 | 0.02 | 0.37 | 2.23 | 0.54 | 0.33 | NO DIF | - | - |
| item2 | 0.03 | | 7.39 | | 0.01 | | Uniform | A | Female |
| item3 | 0.06 | | 25.38 | | 0.00 | | Uniform | A | Female |
| item4 | -0.00 | 0.02 | 0.02 | 3.24 | 0.89 | 0.20 | NO DIF | - | - |
| item5 | -0.03 | | 8.62 | | 0.00 | | Uniform | A | Male |
| item6 | 0.05 | | 18.61 | | 0.00 | | Uniform | A | Female |
| item7 | 0.06 | | 30.44 | | 0.00 | | Uniform | A | Female |
| item8 | 0.02 | 0.02 | 2.84 | 2.84 | 0.09 | 0.09 | NO DIF | - | - |
| item9 | 0.05 | | 17.31 | | 0.00 | | Uniform | A | Female |
| item10 | 0.04 | | 10.34 | | 0.00 | | Uniform | A | Female |
| **item11** | **0.10** | | **82.43** | | **0.00** | | **Uniform** | **C** | **Female** |
| item12 | 0.02 | 0.02 | 2.56 | 2.56 | 0.11 | 0.11 | NO DIF | - | - |
| item13 | 0.00 | 0.00 | 0.08 | 0.08 | 0.77 | 0.77 | NO DIF | - | - |
| item14 | 0.02 | 0.02 | 3.60 | 3.60 | 0.06 | 0.06 | NO DIF | - | - |
| item15 | 0.03 | | 7.72 | | 0.01 | | Uniform | A | Female |
| item16 | 0.05 | | 11.64 | | 0.00 | | Uniform | A | Female |
| item17 | -0.02 | 0.03 | 3.47 | 5.45 | 0.06 | 0.07 | NO DIF | - | - |
| item18 | -0.02 | | 3.95 | | 0.05 | | Uniform | A | Male |
| item19 | 0.02 | 0.02 | 2.82 | 2.82 | 0.09 | 0.09 | NO DIF | - | - |
| item20 | -0.00 | 0.02 | 0.01 | 3.13 | 0.94 | 0.21 | NO DIF | - | - |

Table 9 shows the results of DIF analysis when item purification was performed in Lord's $\chi^2$ and Raju's unsigned area methods.

**Table 9.** *DIF analysis results based on Lord's $\chi^2$ ve Raju's unsigned area when item purification was performed.*

| Item | Lord's $\chi^2$ | *p*-value | Raju's Z | *p*-value | item | Lord's $\chi^2$ | *p*-value | Raju's Z | *p*-value |
|---|---|---|---|---|---|---|---|---|---|
| item1 | 2.80 | 0.25 | -1.74 | 0.08 | **item11** | **31.49** | **0.00** | **-4.36** | **0.00** |
| item2 | 42.31 | 0.00 | -5.95 | 0.00 | item12 | 0.14 | 0.93 | -0.56 | 0.57 |
| item3 | 1.49 | 0.47 | -1.06 | 0.29 | item13 | 28.60 | 0.00 | 3.84 | 0.00 |
| item4 | 6.03 | 0.05 | -2.73 | 0.01 | item14 | 4.14 | 0.13 | -1.85 | 0.06 |
| item5 | 3.16 | 0.21 | -1.65 | 0.10 | item15 | 0.92 | 0.63 | -0.94 | 0.35 |
| item6 | 25.08 | 0.00 | -4.42 | 0.00 | item16 | 23.69 | 0.00 | 4.66 | 0.00 |
| item7 | 1.06 | 0.59 | 0.84 | 0.40 | item17 | 1.33 | 0.51 | -1.52 | 0.13 |
| item8 | 3.85 | 0.15 | -1.82 | 0.07 | item18 | 16.53 | 0.00 | -4.40 | 0.00 |
| item9 | 14.56 | 0.00 | 3.64 | 0.00 | item19 | 18.61 | 0.00 | -4.29 | 0.00 |
| item10 | 0.27 | 0.87 | 0.43 | 0.67 | item20 | 5.10 | 0.08 | 2.20 | 0.03 |

Table 9 shows that when item purification was performed, DIF based on both Lord's $\chi^2$ and Raju's unsigned area methods was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, item 18, and item 19. Item 20 only contained DIF based on Raju's unsigned area method.

### 3.3. Comparison of DIF Results

Table 10 contained items for which DIF was detected in different conditions.

**Table 10.** *Comparison of DIF methods.*

| Mantel-Haenszel test | | SIBTEST | | IRT Lord's $\chi^2$ | | IRT Raju's area test | |
|---|---|---|---|---|---|---|---|
| Without item purification | With item purification | Without item purification | With item purification | Without item purification | With item purification | Without item purification | With item purification |
| item1, item2, item3, item4, item5, item6, item7, item11, item13, item16, item17, item18, item20 | item1, item3, item4, item5, item7, item11, item12, item13, item17, item18, item20 | item1, item2, item3, item4, item5, item6, item7, item9, item11, item16, item17, item18, item20 | item2, item3, item5, item6, item7, item9, item10, item11, item15, item16, item18 | item2, item4, item6, item9, item11, item13, item16, item17, item18, item20 | item2, item4, item6, item9, item11, item13, item16, item18, item19 | item2, item4, item6, item9, item11, item13, item16, item17, item18, item19, item20 | item2, item4, item6, item9, item11, item13, item16, item18, item19, item20 |
| 13 items | 11 items | 13 items | 11 items | 10 items | 9 items | 11 items | 10 items |

It was found that the results of DIF analysis differed partially with or without item purification. For example, item 20 contained DIF based on four methods when item purification was not performed, while it contained DIF based on the MH and Raju's unsigned area method when item purification was performed. When item purification was not performed, DIF was detected in item 17 based on four methods, whereas when item purification was performed, it contained

DIF only based on the MH method. Whether item purification was performed or not, DIF was detected in item 3, item 5, item 7, item 11, and item 18 based on SIBTEST and MH methods. Whether item purification was performed or not, DIF was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, and item 18 based on the Lord's chi-square method. Based on Raju’s unsigned area test, DIF was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, item 18, item 19, and item 20 in both cases. When item purification was not performed, DIF was detected in 13 items based on SIBTEST and MH methods. Based on the Lord's chi-square method, DIF was detected in 10 items and 11 items based on Raju's unsigned area method. When item purification was performed, DIF was detected in 11 items based on the SIBTEST and MH methods. Based on Lord's chi-square method, DIF was detected in 9 items. Based on Raju's unsigned area method, DIF was detected in 10 items. It was found that the results differed partially from method to method and according to the condition of item purification.

For the MH method, the average pairwise percent agreement was found to be 80% whether or not purification was performed. For the SIBTEST method, the average pairwise percent agreement was found to be 70% in the same condition. For the Lord's chi-square method, the agreement was 85%, and for Raju's unsigned area method, it was 95% in the same condition. The highest level of agreement was found for Raju's unsigned area method in the condition where purification was both performed and not performed.

### 3.4. Expert Opinions for Bias Study

Table 11 summarizes the opinions of the experts on item 11, for which at least a moderate DIF level was detected as a result of the DIF analysis. Looking at Table 11, it can be seen that there were 6 experts (85.71%) who stated that item 11 was not biased, while one expert stated that it was biased. In the face-to-face interviews with Expert 2, she expressed that she was not sure about the bias of the item.

**Table 11.** *Expert opinions for the item 11.*

| Expert number | Decisions | Expert number | Decisions |
|---|---|---|---|
| E1 | No bias | E5 | No bias |
| E2 | Bias | E6 | No bias |
| E3 | No bias | E7 | No bias |
| E4 | No bias | Total | 6 no bias (%85,71), 1 bias (%14,19) |

Expert 2’ explanation was as follows: "Expression types and patterns in the content of the visual and item used together to cause a bias towards gender… It is known that reading texts and writing activities vary according to gender in terms of preferences. The interests of male and female students may differ depending on the genre. For example, in the case of reading activities, it has been found in the literature that "males prefer to read texts for a purpose such as gaining knowledge and learning how to do something” or, in written activities, “females are much more interested in the activity of writing a letter to a pen pal than male students” ... As a genre in which feelings and thoughts are conveyed, the letter is a communication and bonding tool for female students. From the first years of education, it can be observed that girls write to each other in short notes or in letter format. For these reasons, the expression tested in the question creates more familiarity for girls as a type of writing and can make it easier for them to notice the details in the image. On the other hand, the fact that boys have more access to technological communication tools in terms of opportunities causes them to spend more time with such tools. Therefore, the expressions in e-mails, voicemails, and text messages in the options may attract their attention more as a distraction and may cause them to turn to a wrong answer without fully evaluating the image."

Expert 4 (E4), on the other hand, expressed a different opinion and said, "When the root of the item in the test and the options were examined, no situation that could cause a bias in terms of gender was observed. ". When we asked, “What is your opinion about the item containing DIF in favor of girls/DIF source?” questions, he said “The question is a question that can be answered correctly according to the detail in the image. The fact that female students are more successful in recognizing details than male students may be a source of DIF. And this points to the real difference in students' ability levels, not bias.”. When asked about the reason for this situation, he said, “The question is a question where the correct answer can be found according to the detail in the image. The fact that female students are more successful in recognizing details than male students can be a source of DIF”.

Similarly, Expert 5 (E5) said: "I think that the question does not include item bias that causes female students to be advantageous… Possible sources that may cause DIF are not effective in this question. The act of writing a letter is not closer to female students or further away from male students in terms of cognitive, cultural, curriculum content, or socio-economic terms. In the image given regarding the question, a situation that provides an advantage to female students was not evaluated."

When the opinions of the other experts were examined, it was found that they made statements similar to those of experts 4 and 5. Considering that Expert 2 also gave an undecided opinion, it was found that Item 11 was not biased.

## 4. DISCUSSION and CONCLUSION

This study aimed to detect biased items in the TPSEE which was administered in the English subtest in April 2017. Since methods based on IRT were considered first, unidimensionality and local independence from IRT assumptions were tested. After providing the assumptions, the data fit of the model was examined according to 1PL, 2PL, and 3PL. While the best fit was found to be 3PL by all criteria, the first step was to determine if DIF with and without item purification was performed and when it was not performed based on MH, SIBTEST, Lord's $\chi^2$, and Raju's unsigned area methods. In the analysis of IRT-based methods, DIF analyses were performed based on the 3PL model. When item purification was not performed based on at least two methods, only one item (item 11) was found to contain moderate (B level) DIF. When item purification was performed, DIF was detected at C level based on SIBTEST method and at B level according to MH in item 11. In the conceptual process-based bias analysis, the opinions of seven experts were sought on item 11 and six experts stated that the item was not biased. The reason for the DIF in item 11, where at least moderate DIF was found in three conditions, was that the females were more knowledgeable and had more vocabulary than the experts who were consulted. This situation indicated real differences, not bias.

When conducting DIF analysis using the four methods, it was found that the overall results were generally consistent but somewhat divergent. DIF analysis results varied with or without item purification. In the absence of item purification, DIF was detected in the same number of items based on the SIBTEST and MH methods. The MH and SIBTEST methods produced partially similar results. Compared to the Lord's chi-square method and the Raju's unsigned area test methods, DIF was detected in fewer items. When item purification was performed, the total number of DIF-containing items determined based on the four methods decreased. Except for item 11 all the items were unbiased items. One expert identified bias in item 11 but stated that she was not sure when deciding that during the face-to-face interviews.

The research findings indicate that the research results may differ based on the method used. Examination of DIF studies in the literature shows that the results may differ based on the methods (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019; Çepni & Kelecioğlu, 2021; Soysal &

Yılmaz Koğar, 2021). Camilli and Shepard (1994) suggest that DIF detection should be based on at least two methods. Thus, more accurate inferences can be made by comparing the results obtained from methods with different statistical backgrounds.

While there were studies in the literature that reveal the effect of purification on DIF detection (e.g., Özdemir, 2015; Tunc et al., 2018; Soysal & Yılmaz Koğar, 2021), it was seen that no bias study was conducted in any of these studies. Therefore, further bias studies are required. In his research on DIF detection using Lord's Chi-Square, Raju's Area and Likelihood-Ratio Test methods in the 2011 TIMSS mathematics subtest, Özdemir (2015) found that performing purification caused a difference in the number of items in which DIF was detected, especially in Lord's Chi-square and Raju's Area methods. When the item was purified, the number of DIF-detected items in these two methods decreased in this study, which is consistent with our study. However, purification or non-purification in the Likelihood-Ratio Test method did not cause such a difference. Soysal and Yılmaz Koğar (2021) found that DIF was detected in TPSEE 2016 Turkish subtest using Raju's unsigned area and Lord's $\chi^2$ methods, and DIF was detected in more items when item purification was performed. Similarly, Tunc et al. (2018) reported the same findings as Soysal and Yılmaz Koğar's (2021) research; however, their results differ from those of our research.

When the studies in the literature were analyzed, it was seen that while there were numerous DIF detection studies, a limited number of studies on bias had been conducted (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019). Upon analysis of all these aforementioned studies, it was revealed that a bias study had been conducted for other courses' tests other than the English language test (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018). In the literature, it was seen that the study of DIF and bias regarding the English language test was quite limited (e.g., Akcan & Kabasakal, 2019). Akcan and Kabasakal (2019) analyzed the items of the English test items of the "Undergraduate Placement Exam (UPE)" administered in 2016 by gender based on "MH, SIBTEST, and Multiple Indicator and Multiple Causes (MIMIC) methods". Their analysis of 60 items revealed that one item in the translation subtest was found to be DIF in favor of male students. Based upon expert opinion, they concluded that this item did not show bias. Using different DIF detection methods led to partially different conclusions in terms of the number of items with DIF and the level of DIF.

The research has several limitations and suggestions. Firstly, there were four methods utilized in the research. Therefore, similar studies could be performed using other DIF detection methods. TPSEE was administered in April 2017 and the "booklet A" dataset was used in the research. Similar studies can be performed on diverse datasets. As binary (1-0) data were used in the research, DIF detection can be conducted in polytomous data. Only gender-based bias has been addressed in recent research. Future researchers can focus on different sources of DIF. The study consulted the opinions of seven experts when determining the biased item. Similar studies can be designed by holding an item bias expert panel with the Delphi Technique. In the study, IRT-based DIF determination based on the 3PL model was performed. The results can be compared by making IRT-based DIF estimations based on 1PL, 2PL, and 3PL. No correction method was applied in this research. The effect of different correction methods on DIF detection can be investigated. This research was carried out based on real data, while simulation studies can be conducted under different conditions.

## Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

## Orcid

Serap Büyükkıdık https://orcid.org/0000-0003-4335-2949

## REFERENCES

Akcan, R., & Atalay Kabasakal, K.A. (2019). An investigation of item bias of English test: The case of 2016 year undergraduate placement exam in Turkey. *International Journal of Assessment Tools in Education*, *6*(1), 48-62. https://doi.org/10.21449/ijate.508581

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

Bakan Kalaycıoğlu, D. (2022). Gender-based differential item functioning analysis of the medical specialization education entrance examination. *Journal of Measurement and Evaluation in Education and Psychology*, *13*(1), 1-13. https://doi.org/10.21031/epod.998592

Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Item bias analysis of the university entrance examination. *Education and Science*, *36*(161), 3–13.

Camilli, G. & Shepard, A.L. (1994). *Methods for identifying biased test items* (1st ed.). Sage.

Chalmers, R.P. (2018). Improving the crossing-SIBTEST statistic for detecting non-uniform DIF. *Psychometrika*, *83*(2), 376-386.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory* (1st ed.). Holt, Rinehart and Winston.

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. https://doi.org/10.21031/epod.988879

Emily, D., Brooks, G., & Johanson, G. (2021). Detecting differential ıtem functioning: Item response theory methods versus the Mantel-Haenszel procedure. *International Journal of Assessment Tools in Education*, *8*(2), 376-393. https://doi.org/10.21449/ijate.730141

Fidalgo, A.M., Mellenbergh, G.J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*(3), 43-53.

Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, *8*(1), 10-16.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313-334. https://doi.org/10.1207/s15324818ame0204_4

Holland, P.W., & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, (2), i-24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Holland, P.W., & Wainer, H. (Eds.) (1993). *Differential item functioning* (1st ed.). Lawrence Erlbaum.

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

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

Khalid, M.N., & Glas, C.A. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186-197. https://doi.org/10.1016/j.measurement.2013.12.019

Li, H.H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677

Llach, M.P.A., & Gallego, M.T. (2012). Vocabulary knowledge development and gender differences in a second language. *Elia*, *12*(1), 45-75.

Lord, F.M. (1980). *Applications of item response theory to practical problems* (1st edition). Erlbaum.

Magis, D., & Facon, B. (2013). Item purification does not always improve DIF detection: A counterexample with Angoff's delta plot. *Educational and Psychological Measurement, 73*(2), 293-311. https://doi.org/10.1177/0013164412451903

Martinkova, P., & Drabinova, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal, 10*(2), 503-515. https://doi.org/10.32614/RJ-2018-074

Osterlind, S.J. (1983). *Test item bias* (1st ed.). Sage.

Özdemir, B. (2015). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia-Social and Behavioral Sciences*, *174*, 2075-2083. https://doi.org/10.1016/j.sbspro.2015.02.004

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502. https://doi.org/10.1007/BF02294403

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. https://doi.org/10.1177/014662169001400208

Roussos, L., & Stout, W. (1996) A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20,* 355-371. https://doi.org/10.1177/014662169602000404

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187. https://doi.org/10.1080/13803611.2013.767621

Soysal, S., & Yılmaz Koğar, E.Y. (2021). An investigation of item position effects by means of IRT-based differential item functioning methods. *International Journal of Assessment Tools in Education, 8*(2), 239-256. https://doi.org/10.21449/ijate.779963

Tunc, E.B., Uluman, M., & Avcu, A. (2018). Revisiting the effect of ıtem purification on differantial ıtem functioning; real data findings. *International Online Journal of Educational Sciences, 10*(5), 139- 147. https://doi.org/10.15345/iojes.2018.05.010

Wang, W.C., & Su, Y.H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113-144. https://doi.org/10.1207/s15324818ame1702_2

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

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*(2), 447-470.

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development*. In P. W. Holland, & H. Wainer, Differential Item Functioning (pp. 337-347). Erlbaum.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters, 160. https://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf