



## An Investigation of Item Position Effects by Means of IRT-Based Differential Item Functioning Methods

Sumeyra Soysal <sup>1,\*</sup>, Esin Yilmaz Kogar <sup>2</sup>

<sup>1</sup>Necmettin Erbakan University, Faculty of Education, Department of Educational Sciences, Konya, Turkey

<sup>2</sup>Niğde Ömer Halisdemir University, Faculty of Education, Department of Educational Sciences, Niğde, Turkey

### ARTICLE HISTORY

Received: Aug. 13, 2020

Revised: Jan. 08, 2021

Accepted: Feb. 11, 2021

### Keywords:

Item position effects,  
Item Response Theory,  
Differential item function,  
Raju's unsigned area,  
Lord's chi-square.

**Abstract:** In this study, whether item position effects lead to DIF in the condition where different test booklets are used was investigated. To do this the methods of Lord's chi-square and Raju's unsigned area with the 3PL model under with and without item purification were used. When the performance of the methods was compared, it was revealed that generally, the method of Lord's chi-square identified more items with DIF than did the method of Raju's unsigned area. The differentiation of the booklets with respect to item position resulted in a higher number of items displaying DIF with item purification conditions. Based on the findings of the present study, to avoid the occurrence of DIF due to item position effects, it is recommended to position the same items across different booklets in similar locations when forming different booklets.

## 1. INTRODUCTION

With the help of measurement tools used in the field of education, various decisions such as passed/failed, successful/unsuccessful were intended to reach about individuals and it is aimed to affect individuals' lives as accurately as possible. Various methods are used in large-scale assessments in education in line with this aim. To make the results of these kinds of assessments more reliable, one of the widely used methods in different positions or locations within the tests (Bulut et al., 2017). Thus, problems such as individuals memorizing items or copying answers of other examinees during the test application can be overcome (Bulut, 2015). Thus, the effect of these factors that may affect the psychometric properties of the test can be reduced. However, although the use of different test forms or booklets has positive aspects, it may lead to psychometric issues such as position effects of items (Bulut, 2015). The consequences of the position effect on individuals' abilities are ignored in many test creation processes. If such an effect occurs, it is assumed to be the same for all persons and all items therefore it is thought to not affect the person's ability or item difficulty (Hahne, 2008). However, in practice, individuals' test scores can vary according to item position (Kleinke, 1980). In that case, item position effects that cause changes in individuals' test scores may threaten the validity of test score interpretations (Trendtel & Robitzsch, 2018). Hence, examining the positioning of the

**CONTACT:** Sümeýra Soysal ✉ [sumeyrasoysal@hotmail.com](mailto:sumeyrasoysal@hotmail.com) 📍 Necmettin Erbakan University, Faculty of Education, Department of Educational Sciences, 42090, Konya, Turkey

same items in various ways across different booklets should be examined and investigated to see whether or not one book type is more advantageous for some groups of test takers which is important for the test development process. The positions of items in booklets or test forms created by item position manipulations may lead to differential item functioning (DIF) (Akayleh, 2018; Balta & Omur Sunbul, 2017; Debeer & Janssen, 2013; Erdem, 2015). The present examines whether item position effects lead to DIF in test items or not.

### **1.1. Item Position Effects**

The interaction between the position of a test item in a test booklet and the performance a test taker displays on the same item is called item position effects – IP effects (Qian, 2014). Kingston and Dorans (1984) stated that, in the most classical way, IP effects may emerge in two conditions; namely, items in a measurement instrument that are positioned towards the end may be found easy by test takers owing to practice or learning effect (a positive IP effect) or they can be found difficult owing to fatigue effect (a negative IP effect).

An item displaying IP effects means that the item parameters (e.g., difficulty or discrimination) can vary according to the item's position in the booklet (Weirich et al., 2017). For example, Weirich et al. (2017) stated that considering IP effects on item difficulty, an item administered at the end of a test often is more difficult than the same item administered at the beginning of the test (p.115). Similarly, Le (2017) concluded that items tend to be more difficult when placed towards the end of the test. The test-takers in this study may have found the items positioned towards the end difficult owing to their decrease in motivation in the exam. However, whatever the underlying reason is, conditions that occur owing to IP effects negatively impact the validity of the results. Various studies have also indicated that it is important to consider position effect to test the validity of an assessment (Hahne, 2008; Hohensinn et al., 2008; Qian, 2014).

Studies in the literature investigated whether creating different test forms, arranging the location of the items in the test, and ordering the items from easy to hard or hard to easy affect the individuals' performance or item parameters. However, the results of the studies that examined this subject are not the same. While some studies have determined that the item position has a role on individuals' performance (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Ollenu & Etsey, 2015; The West African Examinations Council [WAEC], 1993), others have concluded that item position does not affect the performance of students or examinees (Doğan Gül & Çokluk Bökeoğlu, 2018; Perlini et al., 1988; Tal et al., 2008). In some studies, it was determined that the item position caused bias in item parameter estimates (Debeer & Janssen, 2013; Doğan Gül & Çokluk Bökeoğlu, 2018; Hecht et al., 2015; Meyers et al., 2009). Although there is no clear conclusion about the item position on which different studies have been conducted, different booklets are used in many exams for example the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). For the item security in such large-scale assessments (such as memorizing the item by those taking the exam), booklets created with items in different orders and different clusters could be used (Frey, Hartig, & Rupp, 2009). In such test administrations where there is awareness of the possibility of IP effects leading to negative outcomes (such as bias in item parameters, test score differences), booklet design is used as a measure. However, studies are reporting that IP has an impact even in administrations where booklet design is used as a measure (Hartig & Buchholz, 2012; Le, 2007; Martin et al., 2004).

Although the studies on the IP effects are mostly based on Classical Test Theory (CTT), there are also studies conducted with Item Response Theory (IRT) framework, the use of which has become widespread in many fields (Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Qian, 2014; Weirich et al., 2014). The fundamental assumptions of IRT are that the individual's ability measures can be obtained independently of the tests applied to test takers

and that invariant item and ability parameters can be reached (Hambleton et al., 1991). However, this assumption of item parameter invariance could be in the booklets in which the same items are positioned differently in an achievement test (Weirich et al., 2017).

Since IP effects are not the same for every test-takers, ignoring this effect limits to make a fair comparison. Recent research shows that there can be individual differences as a result of IP effects (Debeer & Janssen, 2013; Verguts & De Boeck; 2000). So, this situation may lead to biased ability parameter estimates. Moreover, IP effects can cause a different source of variation which can have an impact on test scores (Tippets & Benson, 1989). For this reason, the IP effects can cause significant validity issues.

IP effects have a crucial role in almost all moderate to extensive lengths tests using different booklets (Leary & Dorans, 1985). And IP effects is a practical concern in the professional development of test instruments in large-scale assessments (Qian, 2014). Therefore, it is highly worthwhile for test developers to focus and to attention on this issue.

## 1.2. Differential Item Functioning

Differential item functioning (DIF) developed by Holland and Thayer (1988) compares the probability of correct answers to items in test takers from different subgroups with the same level of ability. DIF occurs when different groups of the same underlying ability have different probabilities of responding to an item correctly (Holland & Wainer, 1993).

In DIF studies, it is common that there are at least two groups, i.e. focus and reference groups. The focal group generally refers to a minority group or study group, while the majority group is called the reference group (Schmitt & Crone, 1991). However, when naming the groups is not clear, it can be completely random. There are two types of DIF, namely uniform and non-uniform DIF. Uniform DIF exists when an item is constantly in favor of one group over another group across the  $\theta$  continuum (Zumbo, 1999). In other words, almost all members of a group show better performance than almost all the members of the group who are at the same ability levels. Non-uniform DIF occurs when the item provides a relative advantage, the magnitude of which changes as the  $\theta$  level changes, or when a group has a relative advantage at the low  $\theta$  level, whereas the other group has a relative advantage at the high  $\theta$  level (Penfield & Lam, 2000). If an item shows DIF, it does not mean that item is biased. Generally, DIF analysis is considered as the first step in deciding whether an item can be biased towards a particular group. If the factor causing DIF is irrelevant to the construct being measured by the test, it is a source of bias (Karami, 2012). Kamata and Vaughn (2004, p.51) stated that DIF can arise for reasons other than bias, and therefore an item with DIF should be interpreted as "possibly biased item" or simply called "DIF item".

McNamara and Roever (2006, p. 93) have discussed the DIF detecting methods in four categories: (1) Analyses based on item difficulty. These approaches compare item difficulty estimates. (2) Nonparametric approaches. These procedures use contingency tables, chi-square, and odds ratios. (3) Item-response-theory-based approaches which include 1, 2, and 3 parameter logistic models. (4) Other approaches. These include logistic regression, which also employs a model comparison method, as well as generalizability theory and multifaceted measurement, which are less commonly used in classic DIF studies. As IRT methods were employed in the present study, only these methods were focused on. Methods based on IRT essentially compare item parameters or item characteristic curves that show the focus and reference group test-takers' probability of giving correct answers to items (Camilli & Shepard, 1994). The chi-square test and Raju's area measurement, which are used in the present study, are among the most frequently used IRT-based DIF methods.

### **1.3. Differential Item Functioning Based on Position Effects**

There are numerous studies on IP effects on psychometric item characteristics in the related literature (Hambleton, 1968; Hambleton & Traub, 1974; Kelnke, 1980; Klosner & Gellman, 1973; Leary & Dorans, 1985; Lee, 2007; Newman et al., 1988; Perlini et al., 1998). However, there are fewer studies on whether using different forms or booklets in achievement exams leads to certain psychometric problems such as DIF, and in the majority of these studies, while some focus on item order effects by ordering items from easy to difficult, difficult to easy, or randomly based on item difficulty index (Balta & Omur Sunbul, 2017; Çokluk et al., 2016; Freedle & Kostin, 1991; Plake et al., 1988; Ryan & Chiu, 2001), others focus on IP effects (Avcu et al., 2018; Bulut, 2015; Erdem, 2015).

Ryan and Chiu (2001) developed two forms consisting of 40-items which included topics they had addressed, namely algebra, trigonometry, geometry, and analytic geometry. The items in form-1 were ordered from easy to difficult, while the items in form-2 were ordered from easy to difficult based on the topics. This study reported that the variance in item order did not significantly affect the occurrence of DIF. Çokluk, Gül, and Doğan-Gül (2016) administered three different forms in which the items of a 20-item achievement exam in a science and technology course were ordered from easy to difficult, from difficult to easy, and completely randomly to the seventh-grade students. They investigated whether there was DIF in different forms created by positioning items differently via CTT and IRT-based methods. They concluded that positioning items differently caused a significant difference in the probability of the test takers at the same ability level responding correctly to the items.

Another study, conducted by Bulut (2015), aimed to examine the relationship between gender-based DIF and booklet effect stemming from using test booklets in which the same items were used but positioned differently. By using large-scale verbal reasoning test data in the study, Bulut (2015) conducted uniform and nonuniform DIF analyses using CTT-based DIF detection methods. The study revealed that even though the general difficulty level of the booklets for the male and female groups was found to be similar, some items in each test booklet were observed to be marked as showing uniform and non-uniform DIF. In this study, where the number of non-uniform DIF items was found to be higher than the number of uniform DIF items in each type of booklet. It was deduced that different test booklets were problematic in terms of the exam results of male and female test-takers. In another study, conducted by Erdem (2015), whether the subtests of six different courses in the TEOG (Transition System from Elementary Education to Secondary Education) administered during the fall term of the 2014-2015 academic year displayed DIF based on booklet type was examined using CTT based DIF detection methods. The study revealed that, in terms of the test booklet, there was a high number of DIF displaying items in the subtests of Religion, Culture and Ethics, Turkish Revolution History and Kemalism, and Foreign Language (English), while the number of DIF displaying items decreased in subtests of Turkish and Science and Technology. There was no item displaying DIF in the mathematics subtest.

Findings reported by previous studies show that the location and order of items in a test can affect test results. Hence, it can be claimed that the position of test items should be taken into consideration during a test development process. Thus, the present study aimed to examine whether or not IP effects led to DIF arising from using different test booklets. In large-scale assessments in Turkey are not usually administered as a pilot test. Therefore, items cannot be placed in these booklets based on item difficulty indices.

Instead, items addressing similar learning outcomes are generally clustered together and positioned in the booklets based on these clusters. For this reason, IP effects, not item order, is the focus of the present study. Moreover, it was observed that in the studies where IP effects were examined by using data obtained from large-scale exams, mostly CTT based methods

were used to identify DIF. The current study has some strengths since IRT-based DIF methods are used on real data. In IRT-based DIF studies, generally, 1 parameter logistic (PL) or 2PL models are used without checking for model-data compatibility. However, in the present study, the model was selected by testing the model-data fit. It is believed that the results of the present study will provide test developers preparing different booklets with foresight regarding whether IP effects will lead to DIF or not.

## 2. METHOD

The study group of the present study was comprised of 9737 students who took the TEOG exam during the first term in the 8th-grade on 23rd-24th November 2016. The number of male and female participants were 5049 (51.9%) and 4688 (48.1%), respectively.

### 2.1. Instrument

TEOG is a large-scale assessment administered to 8th-grade students by the Ministry of National Education, General Directory of Measurement, Assessment, and Exam Services in Turkey between the years 2013 and 2017. The scores obtained from this exam are used to place primary school graduates in secondary education institutions (Ministry of National Education [MoNE], 2013). TEOG consists of six subtests, each of which includes 20 multiple-choice items. These subtests are (i) Turkish, (ii) Mathematics, (iii) Science and Technology, (iv) Religion, Culture and Ethics, (v) Turkish Revolution History and Kemalism, and (vi) Foreign Languages (English). In this exam, four booklets (A, B, C, D) formed by varying the positions of the same questions were used. In the present study, the data obtained from the TEOG administered during the first term of the 2016-2017 academic year were used. The study focused only on the Turkish subtest.

### 2.2. Data Analysis

In the data analysis phase of the study, first of all, the missing data in the four booklets, each of which included the responses of 2500 students, were deleted. Booklet A was regarded to be the original booklet, and the responses of the students who took Booklet B, C, or D were reorganized according to Booklet A. Finally, the data set was converted to a categorical score of either 0 or 1. The descriptive statistics of the data set by booklet type used in the study are presented in [Table 1](#).

**Table 1.** *Descriptive statistics by booklets.*

Booklet	N	Min	Max	$\bar{X}$	Std. Dev.	Skewness (Std. Error)	Kurtosis (Std. Error)	KR-20
A	2416	.00	20.00	11.082	4.497	.049 (.050)	-.982 (.100)	.816
B	2453	1.00	20.00	10.824	4.525	.084 (.049)	-.912 (.099)	.817
C	2438	1.00	20.00	10.967	4.475	.118 (.050)	-.940 (.099)	.811
D	2430	.00	20.00	11.003	4.427	.083 (.050)	-.927 (.099)	.808
Total	9737	.00	20.00	10.968	4.481	.083 (.025)	-.940 (.050)	.813

There are no clear-cut guidelines for interpreting measures of skewness and kurtosis. However, Huck (2012, p.27) stated that most researchers accept the range between -1 and +1 for approximately normal distribution. When the statistics regarding skewness and kurtosis coefficients in [Table 1](#) are examined, a normal distribution of the data for all the booklets is observed. As the KR-20 reliability coefficients ranged between .81 and .82 across the booklets, the results obtained from these booklets were considered to be reliable. Because values greater than 0.80 are considered to have high reliability (Salvucci et al., 1997).



Whether the data for each booklet are unidimensional or not was examined through a confirmatory factor analysis based on the WLSMV (weighted least squares mean and variance adjusted) estimation method. WLSMV has been recommended for estimating CFA model parameters with categorical variables (Muthén & Muthén, 2010). To run this analysis, the “lavaan” (Rosseel et al., 2019) package in the R software was utilized. The results obtained are summarized in [Table 2](#).

**Table 2.** Dimensionality analysis by booklets.

Goodness of Fit	A	B	C	D	Criterion*
$\chi^2/df$	294.217/170=1.731	336.128/170=1.977	333.534/170=1.961	268.263/170=1.578	$\leq 5$ Moderate fit $\leq 3$ Perfect fit
CFI	.993	.991	.990	.994	$\geq .90$ Good fit $\geq .95$ Perfect fit
NNFI	.992	.990	.989	.993	$\geq .90$ Good fit $\geq .95$ Perfect fit
RMSEA	.017	.020	.020	.015	$\leq .05$ Perfect fit $\leq .08$ Good fit
SRMR	.024	.026	.026	.023	$\leq .05$ Perfect fit $\leq .08$ Good fit

\*Hu & Bentler, 1999; Sümer, 2000; Kline, 2005; Brown, 2006; Hooper, Coughlan & Mullen, 2008.

When [Table 2](#) is examined, the model-data compatibility for each of the four booklets is observed to be a perfect fit. Based on these findings, it was concluded that the measured construct that unidimensional. This outcome also indicates that the data sets displayed local independence (Hambleton et al., 1991). Finally, model-data compatibility analyses were run to decide which unidimensional parametric IRT model was the most appropriate for the data set used in the study. The results that the analyses yielded are summarized in [Table 3](#).

**Table 3.** Comparison of models with the likelihood-based statistics.

Booklet	Model	Model Fit Indices			Difference		
		AIC	BIC	Log-likelihood	$\Delta\chi^2$	$\Delta df$	<i>p</i>
Booklet A (N=2416)	1PL	56918.35	57039.93	-28438.17			
	2PL	56226.22	56457.82	-28073.11	730.1	19	.00
	3PL	55939.61	56287.00	-27909.80	326.6	20	.00
Booklet B (N=2453)	1PL	58145.46	58267.36	-29051.73			
	2PL	57491.92	57724.12	-28705.96	691.5	19	.00
	3PL	57245.04	57593.34	-28562.52	286.9	20	.00
Booklet C (N=2438)	1PL	58016.17	58137.95	-28987.09			
	2PL	57401.83	57633.79	-28660.92	652.3	19	.00
	3PL	57102.99	57450.93	-28491.50	338.8	20	.00
Booklet D (N=2430)	1PL	57598.17	57719.88	-28778.08			
	2PL	57041.56	57273.39	-28480.78	594.6	19	.00
	3PL	56791.39	57139.13	-28335.69	290.2	20	.00
Total (N=9737)	1PL	230673.50	230824.30	-115315.70			
	2PL	228105.80	228393.20	-114012.90	2605.7	19	.00
	3PL	226973.90	227404.90	-113426.90	1172.0	20	.00

When the item parameters obtained from the 1-, 2- and 3PL models and the  $\Delta\chi^2$  differences summarized in Table 3 were examined, it was concluded that the 3PL model is fitted the Turkish subtest of TEOG. For this reason, the 3PL model was used for the DIF analyses run by utilizing the Lord's chi-square (Lord's  $\chi^2$ ) and Raju's unsigned area methods. These two methods were tested for both with and without item purification. Item purification is used to decrease the effect of items displaying DIF based on the results obtained from DIF methods and is, hence, used to increase the validity of the results (Candell & Drasgow, 1988). In IRT-based methods, item purification is realized by rescaling item parameters in both of the two groups generally based on the reference group scale, while in each step of the purification process, all the items identified as DIF are eliminated and the remaining items are rescaled (Magis & Facon, 2012). In the analyses where items with DIF are taken into consideration, there is a high possibility of Type I error occurrence owing to the fact that items without DIF can be identified as items with DIF (Clauser et al., 1993). However, with the item purification approach the inflation in Type I error rates can be avoided and the power to identify items with DIF can be increased (Magis & Facon, 2012). Hence, in the present study, the effect of item purification on DIF results has also been examined. DIF analyses were run with "difR" package in the R software (Magis et al., 2015) and on the maximum likelihood method. The methods used in the research are, in brief, as follows:

### 2.2.1. Lord's chi-square test

Lord's  $\chi^2$  the hypothesis whether the item parameters (depending on the IRT model used) in one group are different from those in other groups. This method looks at whether there are significant differences between the two groups with statistics (Price, 2014). Lord's  $\chi^2$  is for the item characteristic curves (ICCs) equality between reference groups and focus groups, and is calculated using the following equation:

$$\chi^2 = (v_{iR} - v_{iF})' \Sigma^{-1} (v_{iR} - v_{iF})$$

where  $(v_{iR} - v_{iF})'$  is a vector of differences in the  $i$ -th item parameter estimations (discrimination, difficulty, and pseudo-guessing) between the focus group and the reference group, while  $\Sigma^{-1}$  is the inverse of the asymptotic variance-covariance matrix for differences in item parameter estimations. Lord's  $\chi^2$  test allows for detecting uniform or non-uniform DIF among two groups by setting an appropriate item response model (Lord, 1980, pp. 217-223). When the estimated  $\chi^2$  for  $i$ -th item is significant at .05 level in the present study, this item is flagged as DIF.

### 2.2.2. Raju's area method

Raju (1988, 1990) enhanced the formulas from the area method originally proposed by Rudner, Geston, and Knight (1980) for calculating the exact area between two item response functions (IRFs) derived from two different groups, and presented two statistical tests, called signed and unsigned area methods, for assessing whether the area between two estimated IRFs is significantly different from zero for the 1-, 2- and 3PL models. According to Raju (1988), the signed area (SA) is referred to as the difference between two item characteristic curves, whereas the unsigned area (UA) is referred to as the distance. The SA is computed from the difference between item difficulty parameters, whereas the UA is calculated from the difference between both difficulty and discrimination parameters. Thus, the SA is about uniform DIF, while the UA is related to the non-uniform DIF. Raju (1988) showed that when the  $c$ -parameters (pseudo-guessing parameter) are unequal, the area between two IRFs was infinite and that infinite procedures for estimating the area between two IRFs with unequal  $c$ -parameter yield misleading results. Raju (1988, 1990) proposed to make equal or fixed  $c$ -parameters for this problem. Therefore,  $c$ -parameters in the focal group were fixed to those in the reference group of the present study. Raju's UA is calculated through the following equation:

$$\text{Raju's UA} = (1 - c) \left| \left( \frac{2(a_2 - a_1)}{Da_1a_2} \right) \ln \left[ 1 + e^{\frac{Da_1a_2(b_2 - b_1)}{a_2 - a_1}} \right] - (b_2 - b_1) \right|$$

where a, b and c are the estimation of item discrimination, difficulty, and pseudo-guessing estimates, respectively.

### 2.2.3. Identify DIF items

To identify DIF items in the present study, each booklet was analyzed using the Lord's  $\chi^2$  and Raju's UA methods with and without purification, separately. Then DIF items were flagged in each booklet. Booklet A was optionally chosen as the reference group and the remaining booklets were used as focus groups in all analyses. The results are presented in such a way that Booklet A was compared against booklets B, C, and D.

## 3. FINDINGS

With Booklet A being used as the reference group, the data obtained through the pairwise comparisons of the booklets based on the methods of Lord's  $\chi^2$  and Raju's UA are summarized in Tables 4, 5, and 6.

**Table 4.** Results of DIF analysis of the booklet A versus booklet B.

Item		Lord's $\chi^2$		Raju's UA	
Position in A	Position in B	Without purification	With purification	Without purification	With purification
1	4	11.94*	13.27*	-1.05	-.78
2	5	3.99	4.63	-1.38	-1.38
3	6	1.02	1.71	.71	.21
4	3	3.34	2.03	-1.69	-1.60
5	2	5.49	5.74	-1.88	-2.41*
6	1	10.50*	10.55*	-2.81*	-3.73*
7	12	.93	2.29	.61	-4.25*
8	13	4.12	5.95	-1.84	-3.66*
9	14	10.21*	7.68	1.69	1.40
10	15	4.54	4.06	-1.83	-2.01*
11	16	1.49	2.28	-1.22	-2.49*
12	17	3.88	3.24	.81	.41
13	11	3.86	5.35	-1.77	-2.88*
14	10	3.85	2.49	-.75	-1.32
15	9	9.44*	11.29*	-.95	-1.56
16	8	7.53	9.08*	1.43	1.05
17	7	7.07	5.87	-2.19*	-2.80*
18	19	3.38	1.93	-1.52	-1.18
19	20	8.16*	6.14	-1.46	-1.77
20	18	1.21	1.50	-1.00	-1.46

\* $p < .05$

As can be observed in Table 4, in the Lord's  $\chi^2$  method, the items displaying DIF without item purification are items 1, 6, 9, 15, and 19, while items displaying DIF with item purification are items 1, 6, 15 and 16. In the Raju's UA method, items displaying DIF without item purification are items 6 and 17, while those with item purification are identified as items 5, 6, 7, 8, 10, 11, 13, and 17.

As can be observed in Table 5, in the Lord's  $\chi^2$  method, the items with DIF for both with and without item purification conditions are items 1, 2, 13, and 16. In the Raju's UA method, items displaying DIF without item purification are items 13 and 16, while those with item purification are identified as items 10, 13, and 16.



**Table 5.** Results of DIF analysis of the booklet A versus booklet C.

Item			Lord $\chi^2$		Raju's UA	
	Position in A	Position in C	Without purification	With purification	Without purification	With purification
1	6	9.19*	9.73*	.52	.41	
2	3	13.31*	14.12*	-1.32	-1.36	
3	2	4.83	5.43	.24	.21	
4	1	3.79	2.85	-1.62	-1.83	
5	4	1.43	2.11	.35	.09	
6	5	6.52	7.26	-1.12	-1.14	
7	13	2.15	2.50	1.41	1.01	
8	11	5.02	4.57	-1.30	-1.38	
9	12	4.80	4.16	1.87	1.93	
10	9	6.22	5.59	-1.91	-2.07*	
11	8	1.66	2.15	-1.02	-1.02	
12	7	1.09	1.53	.47	.35	
13	14	11.27*	11.51*	-2.29*	-2.33*	
14	15	1.14	1.41	-.17	-0.28	
15	10	5.39	6.33	1.13	1.11	
16	19	7.90*	8.79*	2.05*	2.03*	
17	20	3.63	3.82	-0.66	-.89	
18	17	2.18	2.16	-1.44	-1.51	
19	18	2.58	2.89	-1.58	-1.64	
20	16	.35	.74	.48	.33	

\* $p < .05$ **Table 6.** Results of DIF analysis of the booklet A versus booklet D.

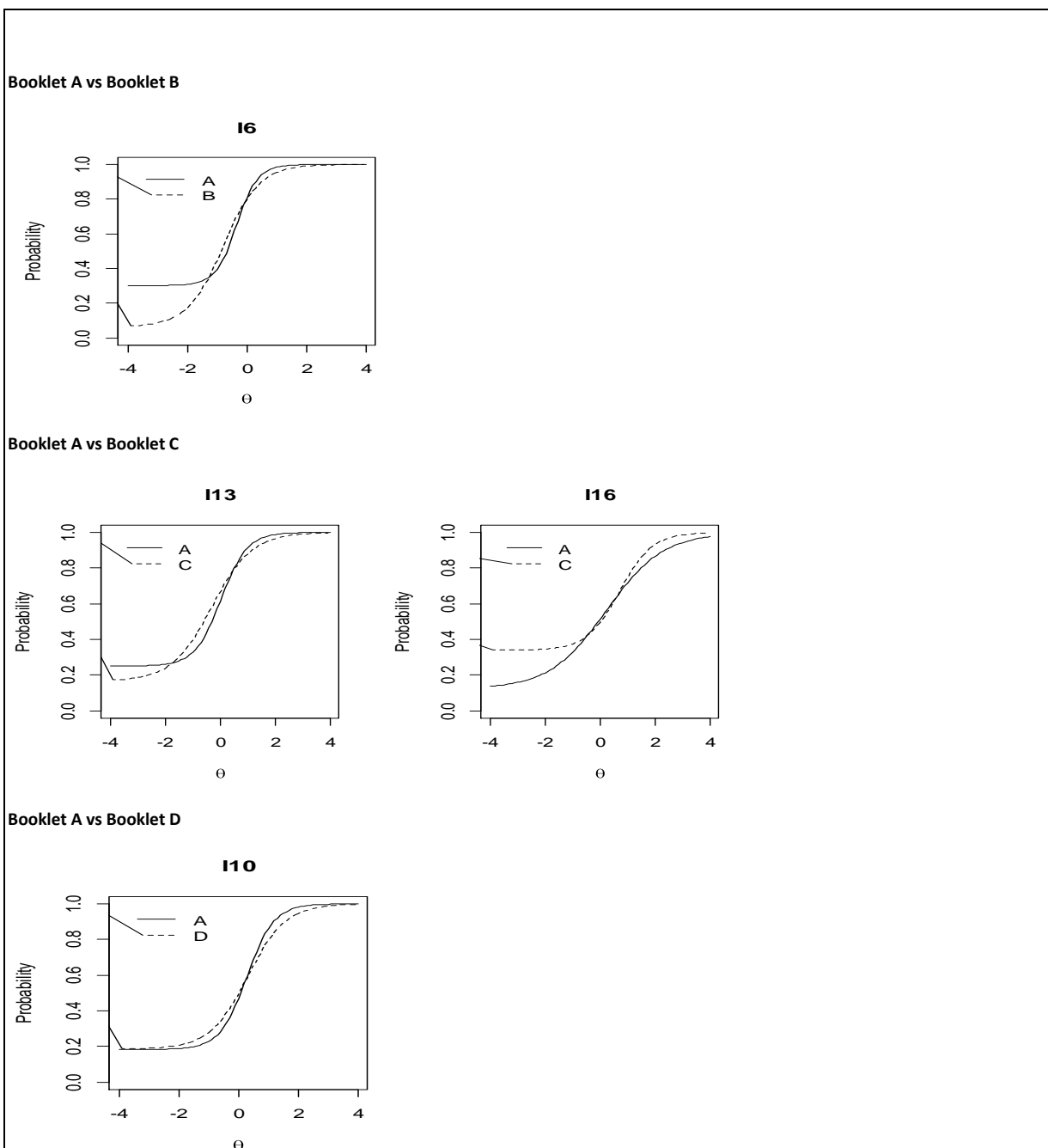
Item			Lord $\chi^2$		Raju's UA	
	Position in A	Position in D	Without purification	With purification	Without purification	With purification
1	2	.13	.30	.22	.13	
2	4	5.77	21.52*	-.28	-1.02	
3	5	4.42	8.98*	-.45	-.67	
4	6	1.18	3.69	-.76	-.42	
5	1	9.86*	15.73*	-1.25	-1.09	
6	3	7.23	23.21*	-1.65	-2.28*	
7	11	4.93	28.85*	-2.04*	-4.14*	
8	14	9.44*	25.83*	-.98	-1.07	
9	13	1.10	1.98	.93	.44	
10	16	10.21*	18.06*	-2.24*	-2.31*	
11	17	3.60	13.14*	-1.78	-2.50*	
12	15	8.70*	7.17	-2.10*	-2.45*	
13	10	2.47	12.23*	-.99	-1.87	
14	9	1.83	5.97	-1.03	-1.29	
15	12	.55	4.04	.50	.30	
16	7	3.60	5.42	1.50	1.24	
17	8	2.76	5.23	-1.48	-2.16*	
18	20	4.00	6.16	-1.53	-1.01	
19	18	3.34	3.76	-1.47	-1.65	
20	19	2.01	3.46	-1.24	-1.36	

\* $p < .05$

As can be observed in Table 6, in the Lord’s  $\chi^2$  method, the items displaying DIF without item purification are items 5, 8, 10, and 12, while items displaying DIF with item purification are items 2, 3, 5, 6, 7, 8, 10, 11, and 13. In the Raju’s UA method, items displaying DIF without item purification are items 7, 10, and 12, while those with item purification are identified as items 6, 7, 10, 11, 12, and 17.

Besides, ICCs were examined for the items flagged as DIF in all conditions (methods x purification) Item 6 was flagged as DIF in both Booklet A and Booklet B. Item 6 in Booklet A is item 1 in Booklet B. In the comparison of Booklet A and Booklet C, items 13 and 16 were flagged as DIF. Items 13 and 16 in Booklet A are items 14 and 10 in Booklet C, respectively. In the comparison of Booklet A and Booklet D, only item 10 was flagged as DIF. This item is item 16 in Booklet D. The ICCs of these four items were shown in Figure 1. It could be observed in Figure 1 that these items displayed non-uniform DIF.

Figure 1. ICCs of DIF items flagged by Lord’s  $\chi^2$  and Raju’s UA methods.



#### 4. DISCUSSION and CONCLUSION

In the present study, the effect of using different booklets formed by changing the position of the same items, which is frequently a preferred practice in large-scale tests, on test-takers' responses was investigated. To this end, four booklets of the Turkish subtest in the 2016 TEOG exam was examined. First, Lord's  $\chi^2$  identified more items with DIF than Raju's unsigned area did in the without item purification condition. Then, items flagged as DIF in the Raju's unsigned area method are generally flagged as DIF in the Lord's  $\chi^2$  method, as well.

In the condition of with item purification, as in the condition of without item purification, fewer items with DIF were observed in the Raju's UA method than in the Lord's  $\chi^2$  method. However, the results that both methods yielded were not revealed to be as consistent as they were in the condition of without item purification. In both methods, the items flagged as DIF when Booklet A was compared against booklets B were more than the items Booklet A was compared against booklets C and D. This could result from the fact that the highest level of similarity in terms of item position was between Booklets A and Booklet C. Thus, it made us think that performing item purification with Lord's  $\chi^2$  and Raju's UA methods tended to be more sensitive than performing without purification. The results of the present study showed consistency with those reported by Özdemir (2015), the study of whom yielded results that were obtained in both with and without item purification using the methods of Lord's  $\chi^2$  and Raju's signed area. Özdemir reported that both Lord's chi-square and Raju's signed area (for 1PL) methods with or without item purification affected both the number of DIF items and DIF items.

In the literature, there are not only studies reporting that item position can have an impact on individuals' performance (Leary & Dorans, 1985; Hambleton, 1968; Wu et al., 2019), but also studies reporting that item position can lead to bias in item parameter estimations (Debeer & Janssen, 2013; Meyers et al., 2009). Meyers et al. (2009), who researched the effect of item position based on IRT, stated that 56% of the variance in item difficulty between the two tests stemmed from the change in the order of the items. Similarly, Debeer and Janssen (2013) reported that in the 2006 PISA reading test, the fact that the item was positioned in a cluster further below the test led to estimations of item difficulty. Taking into consideration that the differentiation in the item parameters reflects onto the ICCs, it can be claimed that this can result in statistically significant results in differential item functioning.

In the present study, the fact that the items flagged as DIF are generally positioned at considerably different places between booklets can indicate that DIF may result from the position of the item in the test. To illustrate, among the items flagged as DIF in at least one method, items 6, 9, 15, and 17 in Booklet A are in the order of 1, 14, 9, and 7 in Booklet B, respectively. Thus, the results obtained in the present study display consistency with those reported in the related literature. However, in the present study, the same items positioned close to each other in different booklets were also revealed to flag as DIF in some conditions (with or without purification) in at least one method (e.g. such items as 2 and 13 in Booklet A are in 3rd and 14th order in Booklet C). In this case, the reason underlying DIF may not be based on item position. It may have arisen due to a type 1 error caused by sampling.

With the consideration of the effects of item position on item difficulty, an item positioned at the end of a test is generally more difficult than the same item positioned at the beginning of the test (Hambleton, 1968; Li et al, 2012; Rose et al., 2019; Weirich et al., 2017). In consistence with the literature, the analyses conducted in the present study also yielded similar results. When the items flagged as DIF were examined in at least one method, item 15 in Booklet A was found to be 9 in Booklet B, and this item was found more difficult by the test takers of Booklet A (see [Appendix-1](#)). This could be attributed to the fatigue effect, mentioned in the study by Davis and Ferdous (2005).

There are also studies reporting that ordering items in a test from easy to difficult has an impact on the probability of giving correct responses to the items (Balta & Omur Sunbul, 2017; Çokluk et al., 2016). In the present study, some items flagged as DIF were evaluated within this scope. To illustrate, the first item in Booklet A, which flagged as DIF, was item 6 and item 5 in Booklets B and C, respectively. When [Appendix-1](#), which presents a summary of the item parameters, is examined, it is observed that this item is the most difficult in the test. Hence, starting a booklet with an easy or difficult item can be an advantage or a disadvantage.

In conclusion, based on the findings of the present study, it can be claimed that the method of Lord's  $\chi^2$  has a higher tendency of flagging items as DIF when compared to the method of Raju's UA. Moreover, it should not be ignored that there may be some prediction error in the DIF results obtained from Raju's UA method since the guessing parameters of the focus group is fixed to the ones of the reference group. As can also be observed in the present study, no method can definitely identify the presence of items flagged as DIF. Even though an item flagged as DIF in any method is no evidence that this item has DIF, it may still require this item to be examined. As a criterion, items flagged as DIF in more than one condition can be examined in detail. When item parameters, the positions of the items, and/or their content are examined carefully, conditions that could be causing DIF can be understood. In the present study which focused on the impact of item position on DIF, it was deduced that an item being positioned at first or last when compared to another booklet could provide an advantage or disadvantage to the test takers.

It is believed that the findings of the present study could provide test developers who prepare different booklets with insight into whether or not IP effects may result in DIF. When forming different booklets, to avoid the occurrence of DIF resulting from IP effects, it is recommended that the same items be positioned in similar locations in the different booklets. The present study is believed to be a significant contribution to the related literature as there is a limited number of studies including DIF analysis based on the 3PL model (Choi et al., 2014; Monahan & Ankenmann, 2010; Uysal et al., 2019; Zwick et al., 1995). In fact, no recent study that tested the Raju's area method based on the 3PL model with real data was encountered in the literature. Hence, in future studies, IP effects based on Raju's area with the 3PL model can be compared with other methods under different conditions. With this kind of simulation study, the results obtained in a condition where there is a fixed c-parameter can be examined. Researchers are recommended to conduct further studies examining the effect of item position together with item order and/or item content on DIF.

### **Declaration of Conflicting Interests and Ethics**

The authors declare 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(s).

### **Authorship contribution statement**

**Sumeyra Soysal:** Investigation, Resources, Software, Visualization, Methodology, Formal Analysis, and Writing - **Esin Yilmaz-Kogar:** Investigation, Resources, Methodology, Formal Analysis, and Writing.

### **ORCID**

Sümeýra SOYSAL  <https://orcid.org/0000-0002-7304-1722>

Esin YILMAZ KOĞAR  <https://orcid.org/0000-0001-6755-9018>

## 5. REFERENCES

- Akayleh, A. S. A. (2018). *Precision of the estimations for some methods of the CTT and IRT as a base to display the differential item functions on the different item ordered test formats.* <https://bit.ly/3aJeFKx>
- Avcu, A., Tunç, E. B., & Uluman, M. (2018). How the order of the items in a booklet affects item functioning: Empirical findings from course level data?. *European Journal of Education Studies*, 4(3), 227-239. <http://doi.org/10.5281/zenodo.1199695>
- Balta, E., & Omur Sunbul, S. (2017). An investigation of ordering test items differently depending on their difficulty level by differential item functioning. *Eurasian Journal of Educational Research*, 72, 23-42. <https://doi.org/doi:10.14689/ejer.2017.72.2>
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
- Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education*, 3(1), 7-16. <https://bit.ly/3cKkhqf>
- Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large-scale Assessments in Education*, 5(1), 8. <http://doi.org/10.1186/s40536-017-0042-x>
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12(3), 253-260. <https://conservancy.umn.edu/bitstream/handle/11299/107645/v12n3p253.pdf?sequence=1>
- Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, 6(4), 269-279. <https://doi.org/10.1207/s15324818ame06042>
- Choi, Y., Alexeev, N., & Cohen, A. (2014). DIF analysis using a mixture 3PL model with a covariate on the TIMSS 2007 mathematics test. In *KAERA Research Forum*, 1(1), 4-14. [http://www.columbia.edu/~ld208/KAERA\\_2014.pdf#page=5](http://www.columbia.edu/~ld208/KAERA_2014.pdf#page=5)
- Çokluk, Ö., Gül, E., & Dogan-Gül, Ç. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences: Theory and Practice*, 16(1), 319-330. <http://dx.doi.org/10.12738/estp.2016.1.0329>
- Davis, J., & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue*. American Institutes for Research. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.847&rep=rep1&type=pdf>
- Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, 50(2), 164-185. [https://ppw.kuleuven.be/okp/\\_pdf/DeBeer2013MIPEW.pdf](https://ppw.kuleuven.be/okp/_pdf/DeBeer2013MIPEW.pdf)
- Doğan Gül, Ç., & Çokluk Bökeoğlu, Ö. (2018). The comparison of academic success of students with low and high anxiety levels in tests varying in item difficulty. *Inonu University Journal of the Faculty of Education*, 19(3), 252-265. <https://doi.org/10.17679/inuefd.341477>
- Erdem, B. (2015). *Ortaöğretime geçişte kullanılan ortak sınavların değişen madde fonksiyonu açısından kitapçık türlerine göre farklı yöntemlerle incelenmesi* [Investigation of Common Exams Used in Transition to High Schools in Terms of Differential Item Functioning Regarding Booklet Types with Different Methods] [Unpublished master dissertation]. Hacettepe University. Ankara.



- Freedle, R., & Kostin, I. (1991). *The prediction of SAT reading comprehension item difficulty for expository prose passages* (ETS Research Report, RR-91-29). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1991.tb01396.x>
- Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28(3), 39-53. <https://doi.org/10.1111/j.1745-3992.2009.00154.x>
- Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, 50(3), 379–390. <https://bit.ly/3aHHyGD>
- Hambleton, R. K. (1968). *The effects of item order and anxiety on test performance and stress*. Paper presented at the meeting of American Educational Research Association. <https://files.eric.ed.gov/fulltext/ED017960.pdf>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, 54(4), 418-431. <https://core.ac.uk/download/pdf/25705605.pdf>
- Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*, 75(6), 1021-1044. <https://doi.org/10.1177/0013164415573311>
- Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, 50, 391-402. <https://bit.ly/39Sb9xY>
- Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, 17(6), 497-509. <https://doi.org/10.1080/13803611.2011.632668>
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.129-143). Erlbaum.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods* 6(1), 53-60. <http://arrow.dit.ie/cgi/viewcontent.cgi?article=1001&context=buschmanart>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Pearson.
- Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal*, 2(2), 49-69. (EJ797693). ERIC. <https://eric.ed.gov/?id=EJ797693>
- Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, 11(2), 59-76. <https://psycnet.apa.org/record/2012-28410-004>
- Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8(2), 147-154. <https://conservancy.umn.edu/bitstream/handle/11299/101880/1/v08n2p147.pdf>

- Kleinke, D. J. (1980). Item order, response location, and examinee sex and handedness on performance on multiple-choice tests. *Journal of Educational Research*, 73(4), 225–229. <https://doi.org/10.1080/00220671.1980.10885240>
- Kline, R. B. (2005). *Principles and practice of structural equation modeling*. The Guilford Press.
- Klosner, N. C., & Gellman, E. K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement*, 33, 413–418. <https://doi.org/10.1177/001316447303300224>
- Le, L. T. (2007, July). *Effects of item positions on their difficulty and discrimination: A study in PISA Science data across test language and countries*. Paper presented at the 72nd Annual Meeting of the Psychometric Society, Tokyo.
- Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55(3), 387–413. <https://doi.org/10.3102/00346543055003387>
- Li, F., Cohen, A., & Shen, L. (2012). Investigating the effect of item position in computer-based tests. *Journal of Educational Measurement*, 49(4), 362–379. <https://doi.org/10.1111/j.1745-3984.2012.00181.x>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- Magis, D., Beland, S., & Raiche, G. (2015). *Package 'difR'* (Version: 5.0). [Computer software manual]. Retrieved May 14, 2018. Retrieved from <https://cran.rproject.org/web/packages/difR/difR.pdf>
- Magis, D., & Facon, B. (2012). 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>
- Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (2004). Item analysis and review. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 224–251). TIMSS & PIRLS International Study Center, Boston College.
- McNamara, T., & C. Roever (2006) *Language testing: The social dimension*. Blackwell.
- Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT based common item equating design. *Applied Measurement in Education*, 22(1), 38–60. <https://doi.org/10.1080/08957340802558342>
- Ministry of National Education [MoNE], (2013). *2013-2014 Eğitim-öğretim yılı ortaöğretimi geçiş ortak sınavları e-klavuzu*. Ankara.
- Monahan, P. O., & Ankenmann, R. D. (2010). Alternative matching scores to control type I error of the Mantel–Haenszel procedure for DIF in dichotomously scored items conforming to 3PL IRT and nonparametric 4PBCB models. *Applied Psychological Measurement*, 34(3), 193–210. <https://doi.org/10.1177/0146621609359283>
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus: Statistical analysis with latent variables user's guide 6.0*. Muthén & Muthén.
- Newman, D. L., Kundert, D. K., Lane Jr, D. S., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education*, 1(1), 89–97. [https://doi.org/10.1207/s15324818ame0101\\_8](https://doi.org/10.1207/s15324818ame0101_8)
- Ollenu, S. N. N., & Etsey, Y. K. A. (2015). The impact of item position in multiple-choice test on student performance at the basic education certificate examination (BECE) level. *Universal Journal of Educational Research*, 3(10), 718–723. <https://doi.org/10.13189/ujer.2015.031009>

- 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>
- Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement Issues and Practice*, 19(3), 5–15. <https://doi.org/10.1111/j.1745-3992.2000.tb00033.x>
- Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie Canadienne*, 39(4), 299-307. <https://doi.org/10.1037/h0086821>
- Plake, B. S., Patience, W. M., & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. *Educational and Psychological Measurement*, 48(4), 885-894. <https://doi.org/10.1177/0013164488484003>
- Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement*, 38(7), 518-534. <https://doi.org/10.1177/0146621614534312>
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4), 495-502. <https://link.springer.com/article/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://conservancy.umn.edu/bitstream/handle/11299/113559/v14n2p197.pdf?sequence=1>
- Rose, N., Nagy, G., Nagengast, B., Frey, A., & Becker, M. (2019). Modeling multiple item context effects with generalized linear mixed models. *Frontiers in Psychology*, 10,248. <https://doi.org/10.3389/fpsyg.2019.00248>
- Rosseel, Y., Jorgensen, T., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., & Scharf, F. (2019). *Package 'lavaan'* (Version: 0.6-5) [Computer software manual]. <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf>
- Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, 5, 213-233. <https://doi.org/10.2307/1164965>
- Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, 14(1), 73-90. [https://doi.org/10.1207/S15324818AME1401\\_06](https://doi.org/10.1207/S15324818AME1401_06)
- Salvucci, S., Walter, E., Conley, V., Fink, S., & Saba, M. (1997). *Measurement error studies at the National Center for Education Statistics (NCES)*. U.S. Department of Education.
- Schmitt, A. P., & Crone, C. R. (1991). Alternative mathematical aptitude item types: DIF issues. *ETS Research Report Series*, 1991(2), i-22. <https://doi.org/10.1002/j.2333-8504.1991.tb01409.x>
- Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar [Structural Equation Modeling: Basic Concepts and Applications]. *Türk Psikoloji Yazıları*, 3(6), 49-73. <https://psycnet.apa.org/record/2006-04302-005>
- Tal, I. R., Akers, K. G. & Hodge, K. G. (2008). Effect of Paper color and question order on exam performance. *Teaching of Psychology*, 35(1), 26-28. <https://doi.org/10.1080/00986280701818482>
- The West African Examinations Council [WAEC] (1993). *The effects of item position on performance in multiple choice tests*. Research Report, Research Division, WAEC, Lagos.

- Tippets, E., & Benson, J. (1989). The effect of item arrangement on test anxiety. *Applied Measurement in Education*, 2(4), 289-296. [https://doi.org/10.1207/s15324818ame0204\\_2](https://doi.org/10.1207/s15324818ame0204_2)
- Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling*, 60(2), 241-263. <https://bit.ly/3cQWkh5>
- Uysal, I., Ertuna, L., Ertas, F. G., & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology*, 10(2), 133-148. <https://doi.org/10.21031/epod.534312>
- Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, 24, 151-162. <https://doi.org/10.1177/01466210022031589>
- Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, 38, 535-548. <https://doi.org/10.1177/0146621614534955>
- Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, 41(2), 115-129. <https://doi.org/10.1177/0146621616676791>
- Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, 7(5), 1-21. <https://doi.org/10.1186/s40536-019-0073-6>
- 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*. Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement*, 32(4), 341-363. <https://www.jstor.org/stable/1435217>

## 6. APPENDIX

### Appendix-1 Item parameter estimation for booklets.

Item	Booklet A			Booklet B			Booklet C			Booklet D		
	a	b	c	a	b	c	a	b	c	a	b	c
I1	1.14	2.43	.21	1.04	2.18	.23	1.20	2.23	.24	1.23	2.36	.21
I2	2.22	-.23	.14	2.03	-.19	.19	1.87	-.42	.15	2.12	-.24	.20
I3	1.37	-1.48	.05	1.61	-1.05	.33	1.47	-1.33	.28	1.28	-1.16	.40
I4	2.21	.53	.20	1.95	.62	.20	1.80	.60	.19	2.18	.59	.23
I5	2.39	.76	.19	1.64	.78	.18	2.47	.74	.21	2.01	.84	.24
I6	2.84	-.34	.30	1.66	-.78	.06	2.51	-.50	.30	2.01	-.58	.25
I7	1.73	-1.34	.00	1.76	-1.38	.00	1.82	-1.39	.00	1.55	-1.53	.00
I8	1.99	-.59	.13	1.74	-.85	.00	1.51	-.78	.05	1.66	-.60	.20
I9	1.32	.30	.20	1.57	.62	.24	1.69	.64	.29	1.47	.48	.25
I10	2.21	.29	.18	1.88	.36	.20	1.69	.30	.18	1.58	.32	.19
I11	2.27	-.08	.21	1.93	-.22	.15	2.04	-.21	.18	1.72	-.28	.13
I12	.94	.81	.21	1.21	.85	.22	1.00	.73	.23	.65	.15	.00
I13	2.11	.03	.25	1.85	-.20	.16	1.38	-.30	.17	1.92	-.10	.23
I14	1.34	.47	.21	1.19	.50	.16	1.30	.46	.22	1.07	.40	.17
I15	.92	-.19	.01	.72	-.44	.00	1.34	.25	.23	.99	.09	.12
I16	.97	.23	.12	1.35	.65	.33	1.67	.71	.34	1.31	.72	.31
I17	2.00	.77	.12	1.39	.81	.07	1.84	.80	.15	1.63	.67	.07
I18	2.26	.30	.22	2.25	.44	.24	1.82	.20	.18	2.11	.43	.25
I19	3.47	.81	.27	2.79	.88	.22	2.56	.77	.25	2.65	.85	.24
I20	2.10	1.16	.33	1.67	1.08	.30	2.33	1.12	.34	1.50	1.23	.31