ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 200-221

Investigation of Differential Item and Step Functioning Procedures in Polytomus Items*

Yasemin KUZU** Selahattin GELBAL***

Abstract

This study aimed to compare differential item functioning (DIF) and differential step function (DSF) detection methods in polytomous items under various conditions. In this context, the study examined Kazakhstan and Turkey data obtained from the ICT Familiarity Questionnaire in PISA 2018. Mantel test, Liu-Agresti statistics, Cox β, and poly-SIBTEST methods were used for polytomous DIF analysis while Adjacent Category Logistic Regression Model and Cumulative Category Log Odds Ratio methods were used for DSF analysis. This study was carried out by using "differential category combining, focus group sample size, focus group: reference group sample ratio and DIF/DSF detection method". SAS and R software were utilized in the creation of conditions; SIBTEST was used for poly-SIBTEST analysis and DIFAS programs were used for the other methods. Analyses demonstrated that the number of items with large DIF was higher in the small sample according to the polytomous DIF detecting methods. Likewise, the number of steps with large DSF is higher in large samples according to the DSF methods. However, it was found that the methods give more consistent results in large samples. During the steps, the DIF value was lower in the items containing DSF with the opposite sign; therefore, not performing DSF analysis on an item with no DIF may yield erroneous results. Although the differential category combining conditions created within the scope of the research did not have a systematic effect on the results, it was suggested to examine this situation in future studies, considering that the frequency of marking the combined categories differentiated the results.

Keywords: polytomous differential item function, differential step function, adjacent approach, cumulative approach, AC-LOR, CU-LOR

Introduction

Valid measures are needed for test scores to reflect individuals' real scores and for interpretations to display the correct results. Validity, which is an aspect of theory and evidence (American Educational Research Association et al., 2014) that supports interpretations or decisions made based on test scores, is one of the most important features that must exist in measurement tools. Tests should measure all individuals with the same accuracy, regardless of variables unrelated to the measured construct (Sireci & Rios, 2013). It would be misleading to compare different countries or groups with a test that does not mean the same thing for everyone, in other words, when the degree of serving its purpose varies according to groups or countries. In this respect, the property measured by the test items should be invariant according to individuals, groups, and countries. The invariance of the items means that the response probabilities of the items do not change according to the groups with the same characteristics. Item and test bias are the most important threats to validity (Clauser & Mazor, 1998).

*This study is a part of the doctoral thesis prepared by the first author and conducted under the supervision of the second author.

** Research Assistant Dr., Kırşehir Ahi Evran University, Faculty of Education, Kırşehir -Türkiye, yaseminkuzu@yandex.com, ORCID ID: 0000-0003-4301-2645
*** Prof.Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, sgelbal@gmail.com, ORCID ID: 0000-0001-5181-7262

To cite this article:
Kuzu, Y. & Gelbal, S. (2023). Investigation of differential item and step functioning procedures in polytomus items. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(3), 200-221. https://doi.org/10.21031/epod.1221823
Received: 20.12.2022
Accepted: 9.09.2023

Differential Item Functioning

Detecting the biased items in a test should include, first of all, determining whether the items have a DIF. DIF refers to the fact that the probability of answering an item correctly differs between individuals with the same ability level in different subgroups (Embretson & Reise, 2000; Hambleton et al., 1991). Examination of DIF studies in the literature shows that while dichotomous (two-category) items were studied first, in recent years, detecting DIF has been more common on polytomous items as well as dichotomous items with the widespread use of performance-based evaluation. Unlike dichotomous items, DIF can take different forms in polytomous items due to the number of response categories.

Various DIF detection methods are cited in the literature, and these methods are classified in different ways in different sources (Camilli & Shepard, 1994; Zumbo, 2007; Ellis & Raju, 2003). DIF detection in polytomous items is more complex than DIF detection in dichotomous items. Based on the invariance in polytomous items, the form of invariance may differ in score levels. So that, while invariance cannot be achieved at one score level, it can be achieved at other score levels and in cases where invariance cannot be achieved in the item, DIF can be observed in favor of the reference group at one score level and in favor of the focus group at another score level (Penfield et al., 2008).

Mantel test

The Mantel test statistic, an extension of the Mantel-Haenszel (MH) test, was developed to determine the relationship between matched groups on variables at the ordinal scale level (Mantel, 1963). DIF analysis with the Mantel test includes testing the null hypothesis with statistics on the chi-square distribution at one degree of freedom. In this context, equation of the Mantel test analysis is as follows (Zwick et al., 1993):

$$Mantel\ \chi^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k Var(F_k)}$$

The Mantel statistic has a chi-square distribution with one degree of freedom. The rejection of the null hypothesis as a result of this test indicates that the item contains DIF.

Liu Agresti estimator

Although the Liu Agresti estimator is not as common as other MH based methods, it is a recommended method for DIF analysis for polytomous items (Penfield & Algina, 2003). Odds ratios are used in the Liu Agresti estimation.

Cox's β statistic

Cox's β statistic is a mathematically equal but conceptually a different approach to the Mantel test (Cox, 1958) and it assumes that the data come from a decentralized multivariate hypergeometric distribution with β parameter. The β value is calculated as follows (Camilli & Congdon, 1999).

$$\hat{\beta} = \frac{\sum_k \sum_J J(n_{RJk} - \tau_{Jk})}{\sum_k \zeta_k^2}$$

A significant difference in β value from zero means that the item contains DIF.

Poly-SIBTEST

The poly-SIBTEST statistic used for DIF detection in polytomous items is an extension of SIBTEST used in dichotomous items and is a non-parametric model (Chang et al., 1996).

The SIBTEST method presents an effect size (β) that indicates the DIF values as well as the presence of DIF in the item. The estimation of the β effect size, which is defined as the expected group difference in the item thought to have DIF at each valid subtest score level, is defined as follows:

$$\hat{\beta} = \sum_{k=0}^{n_m} p_k(\bar{Y}_{Rk} - \bar{Y}_{Fk})$$

The β effect size index proposed by Roussos and Stout (1996) as the SIBTEST effect size is also used to interpret the poly-SIBTEST DIF index in dichotomous and polytomous items (Henderson, 2001).

DIF detection methods, which are widely used in polytomous items, are based on examining the invariance at the item level (Penfield & Lam, 2000). In approaches such as Mantel's chi-square statistic (Mantel, 1963) and the Generalized Mantel Haenszel (GMH) statistic (Somes, 1986); a single DIF index is given because the general invariance collected at all score levels is measured. In this case, it cannot be determined from which score level DIF originates. Therefore, efforts to identify possible causes of DIF and the item revision process with the contribution of experts are less efficient after the DIF analyses, making it more challenging both in terms of time and economy.

Differential Step Functioning

Differential step functioning is a comprehensive approach used to describe the "between-group difference" in measured properties in a particular step of a polytomous item (Penfield, 2007). Unlike DIF analyses, which give a single statistic for the item, DSF analyses yield as many statistics as the number of steps in the item. So, the differential step functioning can be viewed as a subset of the differential item functioning that focuses on DIF effects in the item.

Evaluation of DSF in a polytomous item begins by dividing the item into J= r-1 step function (where r is the number of score levels in the item). Each step function defines the probability of progressing, or "stepping through", from each score level to a successively higher score level. If there is a difference between the groups in one or more of the step functions of the item, it is concluded that the item exhibits DSF. DSF analysis can be performed by using different approaches. Logistic regression (French & Miller, 1996) and IRT-based approaches such as Graded Response Model (GRM) (Cohen et al., 1993) and the Partial Credit Model (PCM) (Penfield et al., 2008) can be given as examples. In this study, DSF detection was done with the most common DSF methods used in the literature: Adjacent Category Logistic Regression Model (AC-LOR) and Cumulative Category Log Odds Ratio (CU-LOR) methods.

For this purpose, Penfield's (2008) probability ratio approach was used which compared the probability of success of the focus and reference group members with the same observed score at step j. Accordingly, the test takers are divided into score groups according to the raw total scores of a test with possible score values k = 1, 2, 3,…., K. In this context, the ratio of the probability of success of the reference group at step j to the probability of success of the focus group is calculated as follows:

$$\hat{\alpha}_j = \frac{\sum_{k=1}^{K} A_{jk} D_{jk} \,/\, N_{jk}}{\sum_{k=1}^{K} B_{jk} C_{jk} \,/\, N_{jk}}$$

A_{jk}: Number of reference group members who succeeded in step j.

B_{jk}: Number of reference group members who failed in step j.

C_{jk}: Number of focus group members who succeeded in step j.

D_{jk}: Number of focus group members who failed in step j.

N_{jk}: $A_{jk} + B_{jk} + C_{jk} + D_{jk}$

This value is equivalent to the Mantel-Haenszel probability ratio for dichotomous items and each step is considered as a dichotomous item (Gattamorta & Penfield, 2012). The natural logarithm of $\hat{\alpha}_j$is denoted by $\hat{\lambda}_j$. $\hat{\lambda}_j$ with a value of zero means no DSF, a negative $\hat{\lambda}_j$ value means DSF in favor of the focus group, and a positive $\hat{\lambda}_j$ value means that DSF exists in favor of the reference group.

Adjacent Category Approach

When performing DSF analysis on polytomous items, each of the J step functions is defined using the adjacent category approach, which is consistent with Generalized Partial Credit Model (GPCM). Under this approach, j. step function expresses the probability of successfully progressing from the j-1 score level to the j score level.

Cumulative Category Approach

When performing DSF analysis on polytomous items, each of the J step functions is defined using the cumulative category approach, which is consistent with GRM. Under this approach, j. step function indicates the probability of successfully progressing from 0, 1, …, j-1 score level to j, …, J score level. Therefore, in the DSF analysis under the cumulative approach, all scores are taken into account in total, unlike the adjacent category approach. Therefore, it is very important to know the approach used to define the step function in the interpretation of step level parameters.

DSF analyses are an important component of a comprehensive DIF analysis for polytomous items. In recent years, researchers have argued by citing many reasons that each score level should be taken into account instead of a single total score level while examining the invariance form in polytomous items (Gattamorta & Penfield, 2012). One of these reasons is that many omnibus DIF methods such as the poly-SIBTEST and the Standard Mean Difference (SMD) show relatively low power when the DSF effect changes in sign or values in steps of a polytomous item (Penfield & Algina, 2003; Wang & Su, 2004). The second reason is related to the fact that the omnibus DIF methods give a value representative of the DSF aggregated across all steps, and thus large values of DSF at certain steps may be missed if only one step has a large amount of DSF or if the DSF is of opposite sign across the steps. Therefore, calculating the DSF for each step will allow important information to be noticed and taken into account. Finally, with such an approach, it will be possible to understand which score levels are responsible for the violation of invariance, and thus, information about the possible causes of DIF will be obtained.

Examination of the studies in which DIF and DSF analyses are performed in conjunction shows that they are rather limited. The statistics in the studies were undertaken mostly on simulation data, and when real data were used, the focus was usually on the current situation (Akour et al., 2015; Ayodele, 2017; Benítez et al., 2015; Gattamorta & Penfield, 2012; Miller et al., 2010; Penfield, 2007; Penfield et al., 2008; Penfield, 2008; Penfield, 2010). This study aimed to compare the DIF and DSF detection methods in polytomous items by manipulating the conditions on real data, and in line with this purpose, answers were sought to the following questions.

1. Do the DIF values obtained by polytomous DIF methods change based on differential category combining and focus group:reference group (F:R) sample ratios when the focus group sample size is 200 (small)?
2. Do the DIF values obtained by polytomous DIF methods change based on differential category combining and F:R sample ratios when the focus group sample size is 1000 (large)?
3. Do the DSF values obtained by DSF methods change based on differential category combining and F:R sample ratios when the focus group sample size is 200 (small)?
4. Do the DSF values obtained by DSF methods change based on differential category combining and F:R sample ratios when the focus group sample size is 1000 (large)?
5. Do the DIF values obtained by polytomous DIF methods with differential category combination rule and F:R sample ratios differ according to sample size?
6. Do the DSF values obtained by DSF methods with differential category combination rule and F:R sample ratios differ according to sample size?
7. In terms of DSF, how are the similarity rates in classifying the item steps of the methods according to the sample sizes of the focal group?

Methods

This research conducted with correlational survey model compared the polytomous DIF/ DSF detection methods on the items taken from PISA 2018 under various conditions.

Study Group

The sample of the research included items related to the frequency of digital device use at school (IC011) within the scope of the "ICT Familiarity Questionnaire" in PISA 2018, which was used for students from Kazakhstan, Turkey and the United States of America (USA). In selecting the countries, firstly, the country rankings in the field of reading skills (weighted area) were examined according to the results of PISA 2018, and the countries were divided into three groups as low, medium, and high level. Considering the fact that the relevant survey was not applied to all of the countries participating in PISA 2018, two conditions (success and economic level) were taken into account in addition to answering this survey in selecting the countries. Therefore, Kazakhstan (69th), a non-OECD country, was selected from the low-level group, Turkey (40th), an OECD country, was selected from the middle-level group and the USA (13th), an OECD country, was selected from the high-level group. This study included the results obtained from comparing Turkey-Kazakhstan, which better reflect the results, in order to ensure that the text would be concise and more precise. The results of the Turkey-USA comparison are included in Kuzu (2021).

Data Collection

The research data were obtained from the official internet address of the OECD (https://www.oecd.org/pisa/data/2018database) where the PISA 2018 data were announced. In this context, the data of Kazakhstan and Turkey, for which the "ICT Familiarity Questionnaire" was answered within the scope of PISA 2018, was studied. The questionnaire includes items related to digital media and digital devices such as desktop computers, laptops, smartphones. The questionnaire consists of different sections, such as the possibility of accessing digital tools at home/school or the time allotted to digital devices. In this study, 10 items -5-point Likert type- related to the frequency of use of digital devices in school (IC011) were examined. As a result of expert opinions, it was decided that the items measured the same dimension and could be summed. The scores obtained from the questionnaire varied

between 10 and 50; high scores meant that the frequency of using digital devices at school was high while low scores meant that the frequency of using digital devices at school was low. Table 1 presents the descriptive statistics and score category distributions for each item on the basis of countries.

Table 1

Descriptive Statistics and Score Category Distributions for the Items in the Data Collection Tool

							Score Category Distributions (%)				
Item	Country	$\bar{\chi}$	Sd	Kurtosis	Skewness	Item-Total Correlation	1	2	3	4	5
I1	KAZ	2.71	1.43	-1.30	.20	.63	29.7	17.1	20.0	19.0	14.2
	TUR	1.78	1.16	.41	1.27	.43	61.9	13.2	13.1	8.6	3.3
I2	KAZ	2.38	1.31	-.92	.52	.79	35.8	20.6	21.6	13.7	8.2
	TUR	1.39	.74	2.19	1.79	.58	74.5	13.9	9.9	1.8	
I3	KAZ	2.79	1.30	-1.06	.11	.79	21.9	19.5	27.6	19.3	11.6
	TUR	2.20	1.15	-.64	.57	.52	37.3	22.8	25.9	10.3	3.6
I4	KAZ	2.54	1.32	-1.06	.33	.84	30.3	20.1	24.0	16.4	9.2
	TUR	1.56	.88	1.27	1.47	.65	66.0	16.9	13.1	3.5	.5
I5	KAZ	2.22	1.29	-.73	.69	.78	42.6	18.9	19.7	12.0	6.8
	TUR	1.36	.73	2.82	1.97	.64	77.5	11.3	9.4	1.9	
I6	KAZ	2.17	1.27	-.66	.72	.77	43.5	19.1	19.9	11.5	6.0
	TUR	1.34	.69	2.87	1.97	.60	77.4	12.5	8.9	1.3	
I7	KAZ	2.62	1.27	-.99	.24	.82	25.5	21.3	27.2	17.1	8.8
	TUR	1.96	1.13	-.23	.89	.49	48.8	19.9	20.5	8.0	2.8
I8	KAZ	2.49	1.30	-1.00	.38	.84	31.4	20.7	24.0	15.5	8.5
	TUR	1.52	.86	1.98	1.63	.59	67.9	16.7	12.0	2.7	.7
I9	KAZ	2.54	1.30	-1.03	.34	.84	29.6	20.9	24.4	16.2	8.8
	TUR	1.51	.86	2.31	1.69	.64	68.1	17.0	11.4	2.6	.9
I10	KAZ	2.55	1.32	-1.06	.33	.84	29.7	20.7	23.7	16.4	9.5
	TUR	1.78	1.04	.66	1.21	.61	55.4	20.8	16,5	5,0	2,3

According to Table 1, the highest mean for all countries was obtained in item 3 ($\bar{\chi}_{KAZ} = 2.79$, $\bar{\chi}_{TUR} = 2.20$) and the lowest mean for all countries was obtained in item 6 ($\bar{\chi}_{KAZ} = 2.17$, $\bar{\chi}_{TUR} = 1.34$) in the "ICT Familiarity Questionnaire" in the items related to the frequency of using digital devices at school. However, it was found that the item means were mostly above 2 for the Kazakhstan data and below 2 for the Turkey data. In this case, it can be argued that the students who participated in PISA 2018 from Turkey had a low level of digital device use at school. On the other hand, examination of the score category distributions of the items shows that more than half of the data for Turkey was concentrated in the 1st category in the majority of the items, whereas specifically the 4th and 5th categories were marked less. It is noteworthy that, the 5th category was not marked at all in items 2, 5, and 6 and the ratio of students who marked the 5th category in items 4, 7, 8, and 9 was below 1%. When the Kazakhstan data

was examined, it was found that the distribution spread to all category levels. For both countries, the 1st category was marked the most and the 5th category the least.

Dimensionality

Exploratory factor analysis was performed to examine the dimensionality of the scale. The sample size of the country data was determined by the Kaiser-Meyer-Olkin (KMO) coefficient and the distribution of the data was checked with the Bartlett Test of Sphericity. The KMO coefficients for country data ranged between .87-.94. According to Kaiser (1970), the value of KMO takes a value between 0 and 1, and when this value approaches 1, it means that the sample size is suitable for factor analysis. On the other hand, when the results of the Bartlett Test of Sphericity were examined, the chi-square value was found to be statistically significant for all two countries ($\chi^2_{KAZ(45)} = 101454.496$, $\chi^2_{TUR(45)} = 16161.504$; p<,01) and therefore, the data were suitable for exploratory factor analysis. In this context, the results of the exploratory factor analysis are presented in Table 2.

Table 2

Factors Obtained as a result of Exploratory Factor Analysis and Amount of Variance Explained

			Eigenvalue	% of variance
IC011	KAZ	Factor 1	7.01	70.09
	TUR	Factor 1	6.279	62.791

The result of the exploratory factor analysis demonstrated a single component with an eigenvalue above 1 for the Kazakhstan and Turkey data, therefore it was unidimensional. Table 3 presents the results regarding the factor loadings of the items.

Table 3

Factor Loading Values for Items Found via Exploratory Factor Analysis

	Item	Factor Loading	
		KAZ	TUR
IC011	I1	.69	.66
	I2	.83	.79
	I3	.82	.70
	I4	.88	.85
	I5	.83	.86
	I6	.82	.83
	I7	.86	.72
	I8	.88	.84
	I9	.88	.85
	I10	.88	.80

Table 3 presents the factor loading values obtained for the items as a result of the exploratory factor analysis. In general, factor loading values varied between .66 and .88.

Items examined within the scope of the research: Ayodele (2017) developed a 20-item test and analyzed the research questions by manipulating 2 items. In this study, three items were chosen to be interpreted due to the high number of research conditions. Psychometric properties were taken into account in the

selection of the items, and the items with the highest item-total correlation for the two countries were selected because they had the highest representative power in the scale. Table 4 shows that the items with the highest item-total test correlations for both countries were Items 4, 9, and 10. In this context, polytomous DIF and DSF analysis results for Item 4, Item 9, and Item 10 were reported and interpreted. The results of the research are limited to the data, methods and conditions used in the research.

Conditions that were examined in this study

This section presents the conditions manipulated in the research.

Category combining rule. First of all, items that were currently coded in the 5-point scale type (1-5) were coded as (0-4) in accordance with the working principles of the DIFAS 5.0 program (Penfield, 2013). Since the aim was to change the number of item categories, afterwards, the categories were combined. All possible combinations in category combination were taken into account, paying attention to the fact that the combined categories were adjacent (Gelin & Zumbo, 2003; Göçer-Şahin et al., 2016). Table 4 presents the category combining conditions created for the purpose of this research.

Table 4

Category Combination Conditions Created within the Scope of the Research Goal

		Before Recoding	New categories	Explanation
three-category	1st condition (C1)	(1,2)	0	(1 and 2) and (4 and 5) merged.
		3	1	
		(4,5)	2	
	2nd condition (C2)	1	0	(2 and 3) and (4 and 5) merged.
		(2,3)	1	
		(4,5)	2	
	3rd condition (C3)	(1,2)	0	(1 and 2) and (3 and 4) merged.
		(3,4)	1	
		5	2	
four-category	4th condition (C4)	(1,2)	0	(1 and 2) merged.
		3	1	
		4	2	
		5	3	
	5th condition (C5)	1	0	(2 and 3) merged.
		(2,3)	1	
		4	2	
		5	3	
	6th condition (C6)	1	0	(4 and 5) merged.
		2	1	
		3	2	
		(4,5)	3	
	7th condition (C7)	1	0	(3 and 4) merged.
		2	1	
		(3,4)	2	
		5	3	
five-category	8th condition (C8)	1	0	It has been recoded due to the conditions of the DIFAS 5.0 program.
		2	1	
		3	2	
		4	3	
		5	4	

According to Table 4, a total of eight category combination conditions were obtained in the analysis of the data: three for three-category data, four for four-category data, and one for five-category data.

Sample size and focus group-reference group sample ratio. Another condition examined in the study was the focus group sample size. Sample size is very important in DIF studies. If the sample size is too small, it leads to poor parameter estimation, thus no DIF and if the sample is too large, it may cause hypersensitivity in DIF detection (Ayodele, 2017). For this reason, this study aimed to make the right decision via working with different sample sizes. Examination of the studies conducted with polytomous items demonstrated that the studies were performed with data from at least 440 individuals (40 focus group-400 reference group) while the common approach was to use data of 100 to 2000 people (Ankenmann et al., 1999; Elosua & Wells, 2013; Gonzalez-Roma et al., 2006; Meade & Lautenschlager, 2004; Wood, 2011). The focus group sample in this study was addressed had two different sizes: 200 (small) and 1000 (large). However, (focus group): (reference group) sample ratios were examined in three conditions as 2:1, 1:1, and 1:3. In this case, while the sample size of focus group was 200, the sample size of reference group was 100, 200 and 600; while the sample size of focus group was 1000, the sample size of reference group was 500, 1000 and 3000.

Polytomous DIF/ DSF detection methods. Mantel test, Liu Agresti, Cox's β, and poly- SIBTEST were used to determine DIF while AC-LOR and CU-LOR analyses were performed as DSF detection methods.

Data Analysis

Polytomous DIF Analyses

DIFAS 5.0 program (Penfield, 2013) was used for the Mantel test, Liu Agresti estimation, and Cox's β statistics from among polytomous DIF detection methods. First of all, the data were re-coded to start from 0 as the smallest value in accordance with the operating principles of the relevant program (1=0, 2=1, 3=2, 4=3, 5=4) and the total score was used as the matching variable in the analyses. A research design with 8*2*3*4 = 192 cells was created for polytomous DIF analysis including category combining rules (8), focus group sample size (2), focus group: reference group sample ratio (3), and DIF detection method (4). For interpretation of the Mantel test results, the critical value for Type I error probability at the .01 level was accepted as 6.63. On the other hand, while interpreting the Liu Aggresti statistic, the standardized Liu Aggresti Cumulative Common Log-Odds Ratio (LOR Z) value in the analysis outputs was used. If this value is greater than 2 or less than -2, DIF is present in the item. A positive Liu Agresti statistic points to the existence of DIF in favor of the reference group while a negative statistic points to the existence of DIF in favor of the focus group. Another statistic obtained from DIFAS program outputs in this study was Cox's β statistics. If the Cox Z value, which is obtained by dividing the Cox's β table value by its standard error, is greater than 2 or less than -2, DIF is present in the item. If this value is positive, the existence of DIF works in favor of the reference group, and if it is negative, the existence of DIF works in favor of the focus group (Penfield, 2013).

The last DIF analysis was performed with the poly-SIBTEST method for polytomous items. While interpreting the results of the analysis conducted by using the SIBTEST program, the β value was taken into consideration and the values $|\beta| \geq 0.088$ were marked as DIF (C level) (Roussos & Stout, 1996).

DSF Analyses

DIFAS 5.0 program was used for CU-LOR and AC-LOR statistics to determine whether the items had DSF. The calculated DSF values for each step of each item were examined. In this context, the $\hat{\lambda}_j$ values obtained from the analysis outputs were interpreted and the items showing large DSF in the steps were marked separately for both methods. The $|\hat{\lambda}_j| > 0.64$ criterion was taken into account for marking items with large DSF (Penfield, 2007; Penfield et al., 2008).

The findings section presents the results of the polytomous DIF and DSF analyses for the selected items with the help of graphics. Critical values of each method are indicated with dashed lines to facilitate the

interpretation of the graphs. In this context, in addition to the critical values indicated with dashed lines in Cox's β, the Liu Agresti, and poly-SIBTEST, the values above the line in the Mantel test statistic point to large DIF. Similarly, the values outside the critical values presented by the dashed lines for the DSF analyses indicate that the item step exhibits large DSF under the relevant conditions. In the DIF and DSF graphs, the DIF/ DSF level increases as you move away from the critical values.

Results

Findings Related to Research Question 1

Figure 1 presents the results obtained according to polytomous DIF methods (The Mantel test, the Liu-Agresti statistics, Cox's β, poly-SIBTEST) under varying conditions when the focus group sample size was 200 (small).

Figure 1

The change in the DIF values in the items when the focus group sample size was 200.

The examination of the change in the DIF values in the items in Figure 1 showed that Item 4 did not show large DIF for almost all sample size ratios and under all conditions according to the Mantel test, the Liu-Agresti and Cox β methods and DIF values were below critical values. According to the poly-SIBTEST method, while the first two sample size ratios showed negative large DIF in the last conditions, large DIF was not observed in other conditions and with 1:3 sample size ratio. The examination of Item 9 showed that values close to the critical value were obtained in the first two sample size ratios when

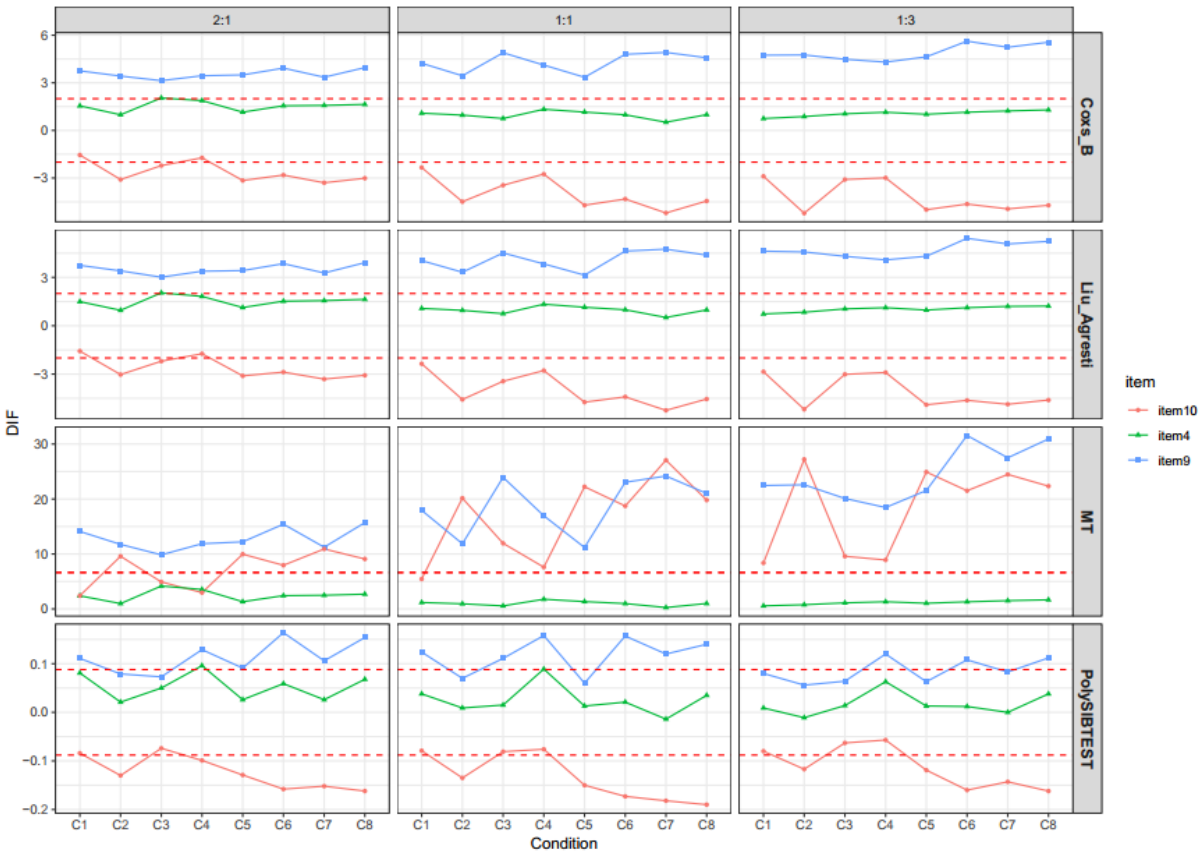
methods other than the poly-SIBTEST method were used. While the sample size ratio was 1:3, there was an increase in the DIF values calculated for the item and a positive large DIF was obtained in all methods and conditions. Finally, the examination of Item 10 showed that the DIF values obtained in the Cox's β and Liu Agresti methods were around the critical value and higher DIF values were obtained at a sample size of 1:3. In the Mantel test and poly-SIBTEST methods, the sudden increase in C5, C6, C7 and C8 conditions was remarkable, especially at the sample size ratio of 1:1. However, these changes did not show a systematic pattern on the basis of conditions.

Findings Related to Research Question 2

Figure 2 presents the results obtained according to polytomous DIF methods (the Mantel test, the Liu-Agresti statistics, Cox β, poly-SIBTEST) under varying conditions when the focus group sample size was 1000 (large).

Figure 2

The change in the DIF values in the items when the focus group sample size was 1000.



The change in the DIF values in the items in Figure 2 was examined. It was observed that Item 4 did not indicate large DIF almost with all sample size ratios and under all conditions for all methods, and DIF values were found to be below the critical values. The examination of Item 9 showed that the first two sample size ratios exhibited a very large DIF in all methods, except for the poly-SIBTEST method. The DIF values obtained were found to increase positively as the sample size ratio increased. In the poly-

SIBTEST method, there were conditions below the critical value as well as large DIF values. Finally, the examination of Item 10 demonstrated that the DIF values obtained in the Cox's β and the Liu Agresti methods were below the critical value in the C1 and C4 conditions at a sample size ratio of 2:1, but exhibited large DIF in all other conditions, with the highest DIF values at the sample size of 1:3. In the poly-SIBTEST method, while large DIF was obtained in some conditions, the values obtained under some conditions were below the critical value. Similar situations were obtained in general based on the sample size ratios.

Findings Related to Sub-Problem 3

Figure 3 presents the results obtained according to the DSF methods (AC-LOR, CU-LOR) under varying conditions when the focus group sample size was 200 (small).

Figure 3

The change in the DSF values in the item steps when the focus group sample size was 200.

When the change in the DSF values in the item steps in Figure 3 was examined, it was seen that large DSF values were obtained in the positive direction in the 1st step of Item 4 under some conditions. The values of DSF obtained from the AC-LOR method were mostly higher than the values obtained from the CU-LOR method. Large DSF values were observed in the negative direction in the other steps of Item 4. Large DSF values were obtained for all conditions and sample ratios in Step 1 of Item 9. The examination of Step 2 showed that the DSF values were below the critical value in all sample ratios and almost all conditions based on the AC-LOR method while positive large DSF was obtained especially in C6, C7, and C8 conditions in the CU-LOR method. While the sample size ratio was 1:2 in Step 3,

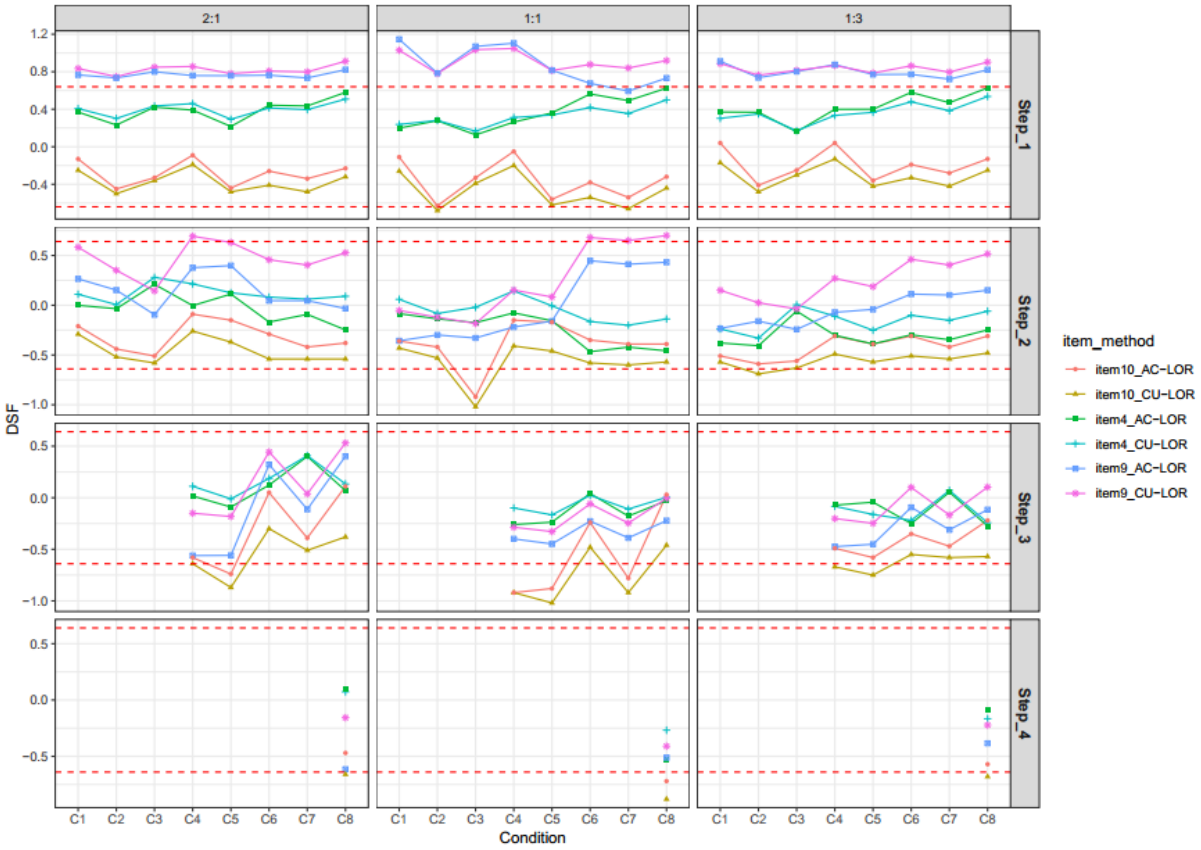
negative DSF was observed in some conditions, while these values remained below the critical value in other sample size ratios. When Step 4 was examined, it was seen that it exhibited higher DSF in the negative direction compared to the AC-LOR method. Finally, the examination of the Steps in Item 10 showed that large DSF was not obtained in the Step 1, except for some conditions where the sample size ratio was 1:2. In Steps 2 and 3, large DSF was obtained mostly in the negative direction. The values of DSF obtained from the CU-LOR method were higher than the values of DSF obtained from the AC-LOR method. Large DSF was observed in all sample size ratios according to the CU-LOR method in Step 4. Large DSF was not obtained with sample size ratios 1:1 and 1:3 with the AC-LOR method.

Findings Related to Research Question 4

Figure 4 presents the results obtained according to DSF methods (AC-LOR, CU-LOR) under varying conditions when the focus group sample size was 1000 (large).

Figure 4

The change in the DSF values in the item steps when the focus group sample size was 1000.



The change in the DSF values in the item steps in Figure 4 was examined and it was seen that large DSF was not observed in the 1st Step of Item 4, except for the C8 condition. DSF values obtained from the AC-LOR method were higher in some conditions while DSF values obtained from the CU-LOR method were higher in other conditions. The DSF values calculated in the other item steps were below the critical value. In Step 1 of Item 9, large DSF values were obtained in almost all conditions and sample ratios. The examination of Step 2 showed that DSF values were below the critical value in all sample ratios and under all conditions in the AC-LOR method; on the other hand, large DSF was obtained in the

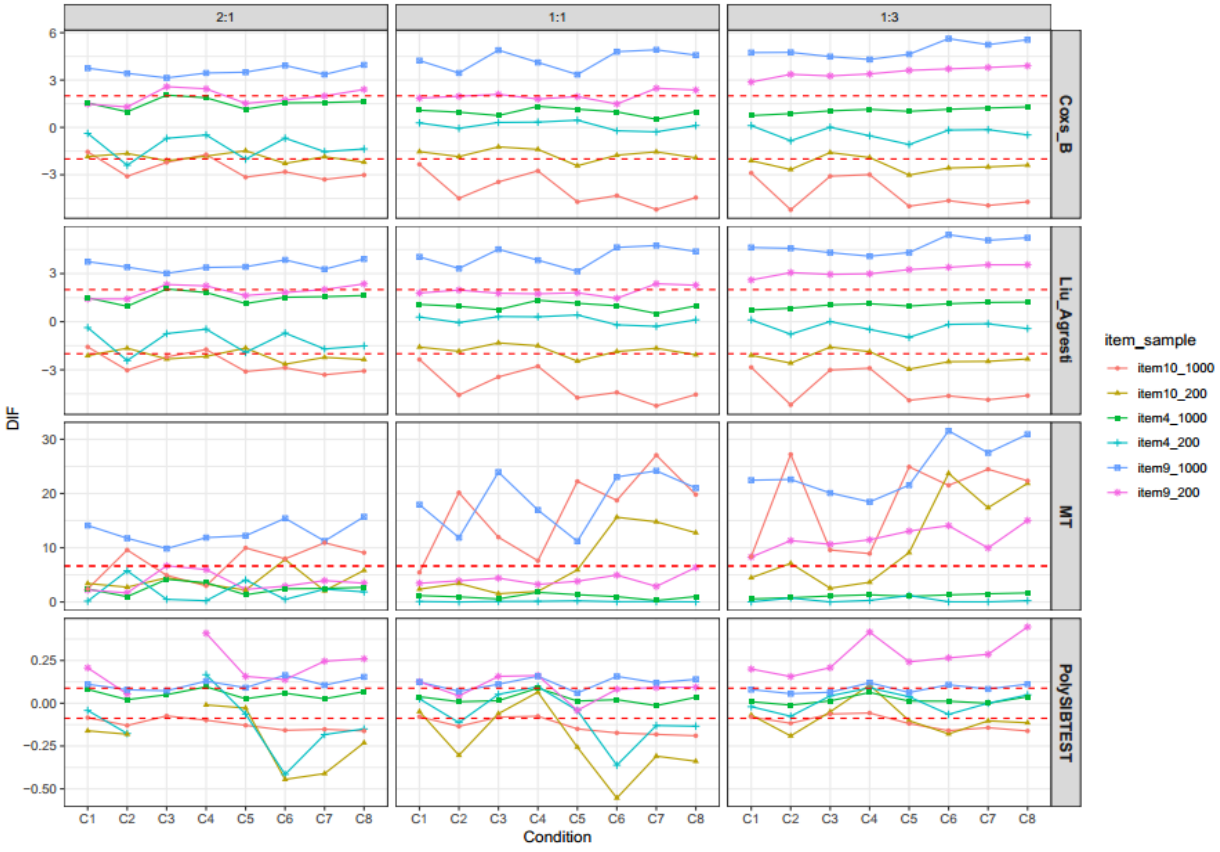
positive direction in some conditions in the CU-LOR method. The DSF values calculated in all conditions and sample size ratios in Steps 3 and 4 were below the critical values. Finally, the examination of the steps in Item 10 demonstrated that large DSF values were not obtained in the 1st Step in general. In Steps 2 and 3, large DSF was obtained in the negative direction under some conditions. The values of DSF obtained from the CU-LOR method were higher than the values of DSF obtained from the AC-LOR method. In step 4, large DSF was observed in all sample size ratios according to the CU-LOR method. Large DSF was not obtained in the AC-LOR method when the sample size ratio was 2:1 and 1:3.

Findings Related to Research Question 5

Figure 5 presents the DIF values obtained by the polytomous DIF methods with differential category combination rule and F: R sample ratios based on the focus group sample size.

Figure 5

The change in the DIF values in items based on focus group sample size



The examination of the change in the DIF values in the items according to the focus group sample size in Figure 5 showed that the DIF values obtained from the large and small samples differed in the opposite direction in item 4, especially according to the Cox's β and the Liu Agresti methods. Accordingly, Item 4 tended to exhibit negative DIF in the small sample, while it exhibited positive DIF in the large sample. When the DIF values related to Item 9 were examined, it was found that the DIF values calculated for both sample sizes were positive, and the DIF values calculated in the large sample were generally large. Unlike other methods, higher DIF values were obtained in the small sample in the poly-SIBTEST method. The DIF values calculated in the small sample at 2:1 and 1:1 sample size ratios were mostly

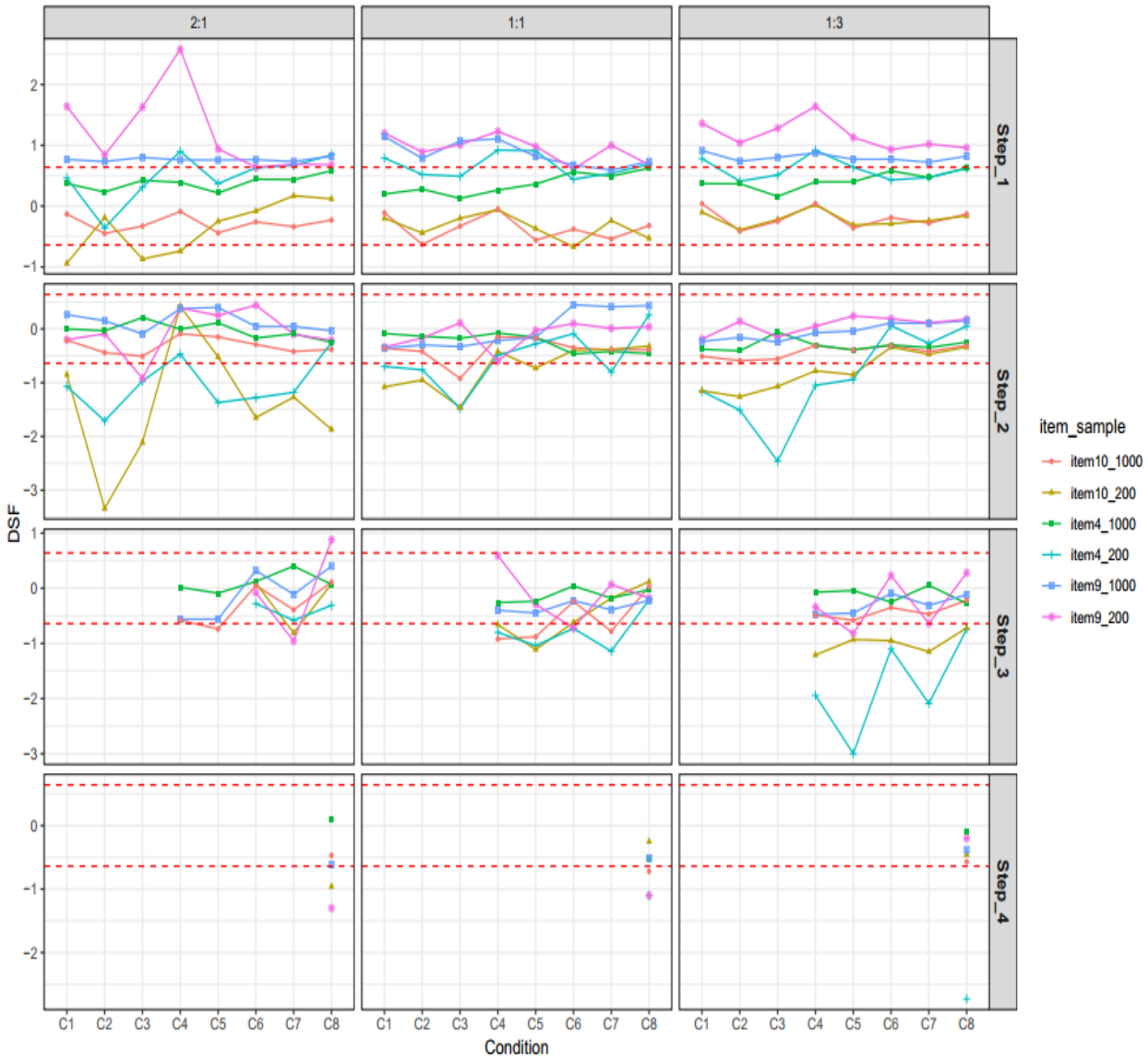
below the critical value in the Cox's β, the Liu Agresti and the MT methods. DIF values calculated on the basis of the conditions did not show a systematic pattern.

Findings Related to Research Question 6

Figure 6(a) presents the DSF values obtained by the AC-LOR method with differential category combination rule and F:R sample ratios based on the focus group sample size.

Figure 6(a)

The change in the DSF values in item steps based on focus group sample size (AC-LOR)

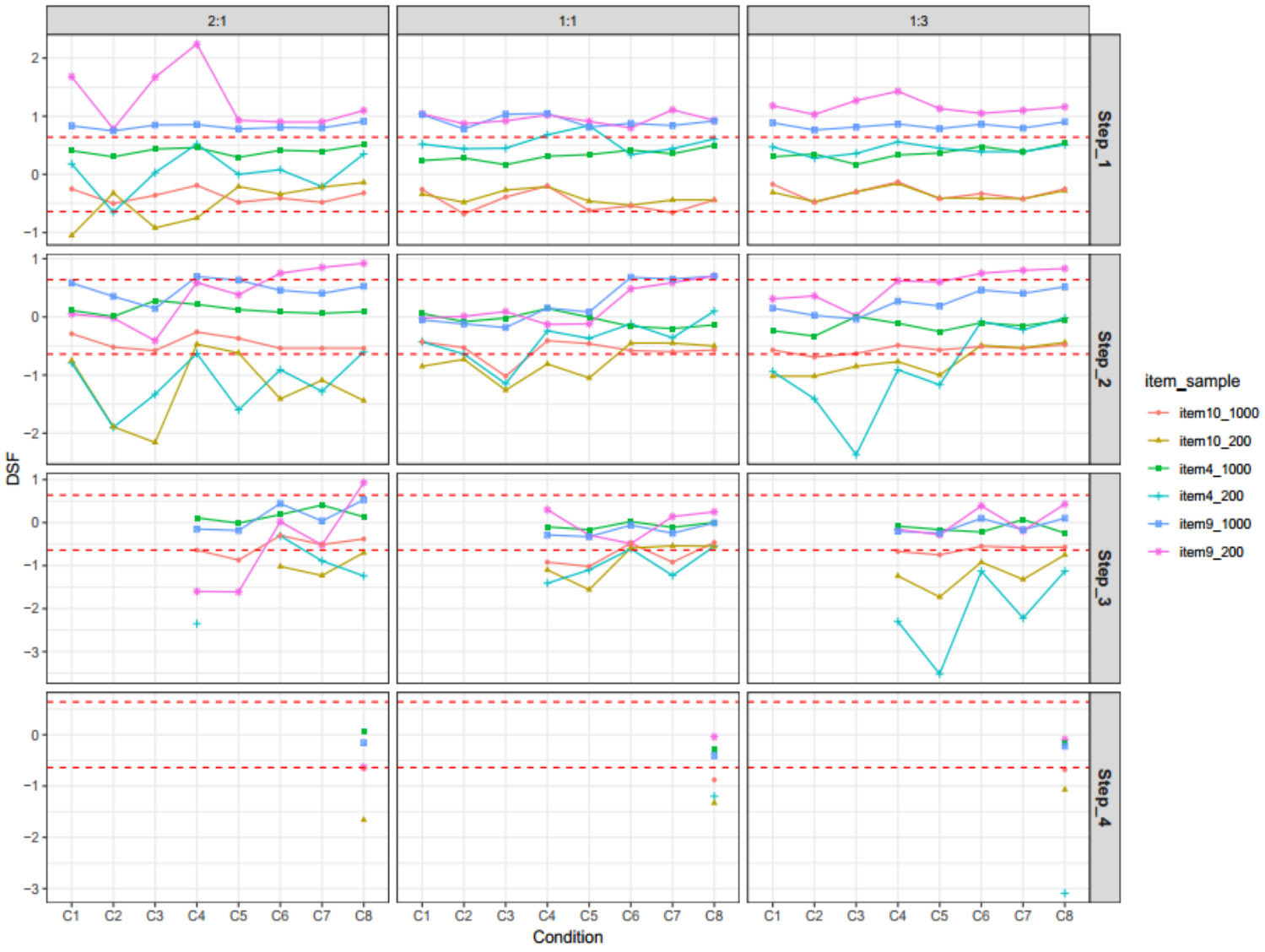


In Figure 6(a), the examination of the change in the DSF values in the item steps according to the focus group sample size based on the AC-LOR method showed that higher DSF was obtained in cases where the focus group sample size was small in Step 1 of Item 4. Similarly, in other steps, large DSF values were obtained when the focus group sample was small. In the steps of Item 9 and Item 10, large DSF was observed in the small sample in general while DSF values were below critical values in the large sample. There was no systematic pattern on the basis of the conditions.

Figure 6(b) presents the DSF values obtained by the CU-LOR method with differential category combination rule and F:R sample ratios based on the focus group sample size.

Figure 6(b)

The change in the DSF values in item steps based on focus group sample size (CU-LOR)



In Figure 6(b), the examination of the change in the DSF values in the item steps according to the focus group sample size based on the CU-LOR method showed that the DSF values calculated in the small and large samples in the 1st and 4th Steps of Item 4 were mostly below the critical value. In the other steps, while large DSF was mostly not observed in the small sample; quite large DSF values in the large sample drew attention. Similar results were obtained in the large and small samples in the steps of Item 9 and Item 10, but slightly higher DSF values were obtained in the small sample. A systematic pattern was not obtained on the basis of the conditions.

Findings Related to Research Question 7

Table 5 presents the similarity ratios of the methods in classifying the item steps in terms of DSF according to the sample sizes.

Table 5

The Similarity Ratios of the Methods in Classifying the Item Steps in Terms of DSF

			Amount of DSF (CU-LOR)							
			Sample size of focal group: 200				Sample size of focal group: 1000			
	Item		Large DSF	Other	Total	Similarity (%)	Large DSF	Other	Total	Similarity (%)
Amount of DSF (AC-LOR)	I4	Large DSF	32	13	45	79.55	-	1	1	98.86
		Other	5	38	43		-	87	87	
		Total	37	51	88		-	88	88	
	I9	Large DSF	32	8	40	77.27	31	-	31	94.32
		Other	12	36	48		5	52	57	
		Total	44	44	88		36	52	88	
	I10	Large DSF	36	1	37	86.36	5	-	5	88.64
		Other	11	40	51		10	73	83	
		Total	47	41	88		15	73	88	

Table 5 provides the number of steps that were marked/unmarked by the AC-LOR and CU-LOR methods as exhibiting large DSF based on sample size. When Turkey-Kazakhstan comparison was evaluated for all items and conditions, it was found that 136 (25.76%) steps showed large DSF based on both methods and it was determined that there was no large DSF compared to both methods in 326 (61.74%) steps. However, in 43 (8.14%) steps, large DSF was detected compared to the CU-LOR method although the AC-LOR method did not mark these steps as large DSF. Likewise, large DSF was calculated according to the AC-LOR method in 23 steps (4.36%) which were marked as without large DSF by the CU-LOR method. The similarity rates in classifying the item steps of the methods in terms of DSF in this comparison changed from 77% to 86% for the focus group with sample size of 200, while they ranged from 89% to 99% for the focus group with sample size of 1000. Therefore, it can be argued that the similarity rates in classifying the item steps of the methods in terms of DSF were higher in the large sample.

Discussion and Conclusion

In this section, the results pertaining to the research problems were discussed in conjunction with the related literature.

Examination of polytomous DIF detection methods (Cox's β, Liu Agresti, MT, and poly-SIBTEST) based on sample size and conditions

The examination of the results obtained from the DIF detection methods shows that the DIF values obtained from Cox's β, the Liu Agresti and MT methods were quite similar to each other in the small sample, while the DIF values obtained from the poly-SIBTEST method differed from the other methods. Among these methods, the poly-SIBTEST helped to detect the highest number of conditions that exhibited large DIF. Although compatible with other methods, the poly-SIBTEST method was found to be the method to detect the items that exhibited the most DIF and provided more sensitive results compared to other methods (Henderson, 2001; Mellor, 1995). It can be argued that the results obtained from the four methods were closest to each other when the sample size ratio was 1:3. However, it was stated that the DIF determination power of the methods tended to decrease with the increase in the sample size of the reference group versus the sample size of the focus group. And It was stated that Type

I error tends to increase in cases where the sample sizes of the reference and focus groups are equal (Wang & Su, 2004; Zwick, 2012)

When the results obtained from DIF detection methods were analyzed in terms of focus group sample size, it was found that all methods provided parallel results when the sample size increased. The highest DIF values and the variability in these values on the basis of the conditions were obtained when the sample size ratios were 1:2 and 1:3. There are studies reporting that the statistical power ratios of the tests are highly affected by the sample size (Bolt, 2002; Kristjansson et al., 2005). Accordingly, it is stated that the methods have a higher statistical power ratio as the sample size increases (Yandı, 2017). When the DIF values obtained from the methods in this study were examined, it was found that the amount of large DIF was higher in the large sample, while the DIF values of the items in the small sample were mostly below the critical values. However, it was observed that the methods provided more consistent results in a large sample.

Examination of DSF detection methods (AC-LOR and CU-LOR) according to sample size and conditions. A comparison of the AC-LOR and CU-LOR methods demonstrated that the DSF values obtained from the AC-LOR method in Steps 1 and 2 for Item 4 were higher than the DSF values obtained from the CU-LOR method. In the other steps of Item 4, the results obtained from the CU-LOR method were found to be higher. On the other hand, the examination of Item 9 demonstrated that the results obtained from the AC-LOR method in some conditions and the CU-LOR method in some conditions were higher in the first two steps, so there was no significant difference between the methods on the basis of the conditions. However, the values of DSF obtained from the CU-LOR method were higher in the other steps of Item 9 and all steps of Item 10. In their study comparing these two methods, Gattamorta and Penfield (2012) stated that there are more steps that exhibit medium to large DSF only according to the effect size in the AC-LOR method used in the adjacent categories approach. When analyzed according to both effect size and significance tests, it was seen that the number of steps exhibiting significant DSF was higher than the CU-LOR method used under the cumulative approach. Due to the smaller standard errors obtained with the CU-LOR method, it was stated that the results were more likely to be statistically significant compared to the AC-LOR method. On the other hand, due to the use of responses from all steps in the cumulative approach, the CU-LOR statistic has higher power than the AC-LOR statistic, which only uses responses in adjacent categories (Ayodele, 2017).

When the DSF detection methods were examined according to sample sizes, it was seen that the DSF values obtained from both methods were higher when the sample was small compared to the large sample. While the same items (Item 4, Item 10) contained half and half DSF in the small sample; they exhibited almost no large DSF in the large sample. On the other hand, the similarity rates in the classification of the item steps of the methods in terms of DSF were higher in the large sample. It clearly shows the importance of the methods used, especially in small samples, when interpreting the invariance and ultimately deciding on the revision or removal of the item.

When the classifications were examined regarding whether the item steps contained large DSF on the basis of methods, it was quite remarkable to note that the similarity rates of the methods were much higher in the large sample. Especially when the sample size was 1000, the percentages of agreement of the methods in the DSF classification made with the CU-LOR and AC-LOR methods of Item 4 and Item 9 were quite high (99% and 94%). Therefore, it can be argued that the methods generated very consistent results, especially in the large sample, in classifying the items in terms of DSF. Parallel to this result, it has been stated in the literature that although the AC-LOR method provides higher DSF values in other DSF classifications, except for small DSF, both methods mostly generate consistent results (Gattamorta, 2009).

When the results of the methods were analyzed on the basis of sample size ratios and conditions, an increase was observed in the DSF values for some items at the same sample size, while a decrease was observed in the DSF values for some items. Therefore, it can be argued that sample size ratios did not have a significant effect on the results of DSF. On the other hand, although the examined conditions did not significantly affect the results, there were fluctuations in the results obtained from the AC-LOR method as the conditions changed. The results show parallelism on the basis of conditions in the CU-

LOR method. In the literature, it is stated that the DSF values estimated under the cumulative approach are more stable than the DSF values estimated under the adjacent categories approach (Gattamorta & Penfield, 2012; Penfield, 2008). It was found that the pattern of the number of steps on the DSF results was not systematic in both methods, whether stable or not. Ayodele (2017) reached similar results and stated that the sample size ratio and the number of steps did not have a statistical and practical significance on the DSF values. Therefore, if the data is polytomous, using the data in its raw form without any changes in the data will produce more valid results. However, if category combining will be used for various reasons, it is recommended to combine categories in accordance with the nature of the research and the data, as which adjacent categories will be combined has no effect.

When the frequency of marking the score categories related to the items was examined, it was observed that approximately half of the individuals concentrated on the first two options in Item 4, Item 9, and Item 10. However, the fact that more than half of the individuals in Turkey data marked the first option made the distribution of categories more skewed. When the creation of the conditions was examined in this context, it was seen that the 1st and 2nd most marked options were combined in conditions 1 and 3 for three-category data and were combined in condition 4 for four-category data. The 4th and 5th least marked options were combined in conditions 1 and 2 for three-category data and in condition 6 for four-category data. DIF analyses showed that the highest DIF values were mostly obtained in condition 2 among conditions 1, 2, and 3 generated for the three-category data. When the four-category data (conditions 4, 5, 6 and 7) were evaluated among themselves, it can be argued that although there was no systematic pattern, more DIF was obtained in condition 6 compared to condition 4. The results of the DSF analysis demonstrated that the results of conditions 1 and 3, in which the first two options were combined in Step 1, differed from the results of condition 2. This differentiation was not systematic and the results of condition 2 were large in some items and small in some others. On the other hand, it can be argued that the DSF values obtained in Step 1 under the conditions created for the four-category data differed between condition 4 and the others. The direction of this differentiation was not standard, while the largest DSF value was obtained in condition 4 for some items, the smallest DSF amount was obtained for some others in condition 4.

Examination of the results obtained from Polytomous DIF and DSF detection methods together

The examination of the studies on DIF and DSF shows that there are studies in which DIF/DSF analyses are performed simultaneously (Akour et al., 2015) or DIF analysis is performed first and then DSF analysis is performed only on DIF-containing items (Miller et al., 2010). Akour et al. (2015) stated that items that do not exhibit large DSF in any of their steps also do not exhibit DIF. However, it has been observed that Type I error is high in some methods that determine DIF when there is no DSF in the item steps (Ayodele, 2017). In other words, although it is rare, cases where a non-DIF-containing item was marked as DIF were encountered in some of the methods. When the results obtained from this study were examined, it was found that Item 4, which did not exhibit DSF at any step in the large sample, was below the critical values of the DIF analysis results, that is, it did not exhibit DIF. On the other hand, when the DSF results for Item 9 were examined when the sample size ratio was 1:1 in the small sample, the DSF values obtained in Steps 1 and 4 were found to be high and with opposite signs. When the DIF results of the related item were examined, it was determined that the item was not DIF according to most of the methods at the same sample size. This may be due to the fact that the DSF values with opposite signs observed in the steps reduce the DIF effect to almost zero. If DSF analysis is not performed on items that do not exhibit DIF, information about the DSF values of the steps cannot be obtained. Therefore, it should be kept in mind that important information about the steps may be overlooked if you first perform the DIF analysis and then perform the DSF analysis only on the DIF-containing items. As a matter of fact, many DIF detection methods have been reported to show relatively low power when the DSF values change in sign and size across steps (Ankenmann et al., 1999; Chang et al., 1996; Penfield & Algina, 2003; Wang & Su, 2004). Therefore, while making decisions for item revision or item removal, it is recommended to perform a DSF analysis on all items, not only on the items with DIF.

When the DIF and DSF analyses were examined together, it was found that in cases where the DIF amount was the highest, the DSF values obtained from the steps of the relevant items varied, but the signs stayed the same.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: This research study complies with research publishing ethics. Secondary data were used in this study. Therefore, ethical approval is not required.

References

Akour, M., Sabah, S., & Hammouri, H. (2015). Net and global differential item functioning in pisa polytomously scored science items. *Journal of Psychoeducational Assessment, 33*(2), 166–176. https://doi.org/10.1177/0734282914541337

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

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*(4), 277–300. https://doi.org/10.1111/j.1745-3984.1999.tb00558.x

Ayodele, A.N. (2017). *Examining power and type 1 error for step and item level tests of invariance: Investigating the effect of the number of item score levels* (Doctoral dissertation). University of Minnesota, USA.

Benítez, I., Padilla, J.L., Hidalgo Montesinos, M. D., & Sireci, S. G. (2015). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*, 29(1), 1–16. https://doi.org/10.1080/08957347.2015.1102915

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113–141. https://doi.org/10.1207/S15324818AME1502_01

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics*, 24(4), 323–341. https://www.jstor.org/stable/pdf/1165366.pdf

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

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomous items: An adaptation of the SIBTEST procedure. *Journal of educational measurement, 33*(3), 333–353. https://doi.org/10.1111/j.1745-3984.1996.tb00496.x

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. An NCME instructional module. *Educational Measurement: Issues and Practice*, 17(1), 31–44. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335–350. https://doi.org/10.1177/014662169301700402

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 20*(2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Ellis, B. B., & Raju, N. S. (2003). Test and Item Bias: What they are, what they aren't, and how to detect them. In J. E. Wall & G. R. Walz (Eds.*), Measuring up: Assessment issues for teachers, counselors, and administrators*. CAPS Press.

Elosua, P., & Wells, C. S. (2013). Detecting DIF in polytomous items using MACS, IRT and ordinal logistic regression. *Psicológica, 34*(2), 327–342. https://www.redalyc.org/pdf/169/16929535011.pdf

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315–332. https://doi.org/10.1111/j.1745-3984.1996.tb00495.x

Gattamorta, K. A. (2009). *A comparison of adjacent categories and cumulative DSF effect estimators* [Doctoral dissertation]. University of Miami, Florida.

Gattamorta, K. A., & Penfield, R. D. (2012). A comparison of adjacent categories and cumulative differential step functioning effect estimators. *Applied Measurement in Education, 25*(2), 142–161. https://doi.org/10.1080/08957347.2012.660387

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, 63*(1), 65–74. https://doi.org/10.1177/0013164402239317

Gonzalez-Roma, V., Hernandez, A., & Gomez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*(1), 29–53. https://doi.org/10.1207/s15327906mbr4101_3

Göçer-Şahin, S., Gelbal, S., & Walker, C. M. (2016, October). *Impact of decreasing category number of polytomous items on DIF* [Conference presentation]. 15th International Mineral Processing Symposium (IMPS 2016), USA.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Henderson, D. L. (2001, April 10-14). *Prevalence of gender DIF in mixed format high school exit examinations*. American Educational Research Association 2001 Annual Meeting, USA. https://files.eric.ed.gov/fulltext/ED458284.pdf

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935–953. https://doi.org/10.1177/0013164405275668

Kuzu, Y. (2021). *Investigation of Differential Item and Step Functioning Procedures in Polytomously Scored Items* [Doctoral dissertation]. Hacettepe University, Ankara.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58(303), 690–700. https://www.jstor.org/stable/pdf/2282717.pdf

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361–388. https://tarjomefa.com/wp-content/uploads/2019/10/F1430-TarjomeFa-English.pdf

Mellor, T. L. (1995). *A comparison of four differantial item functioning methods for polytomously scored items* [Unpublished doctoral dissertation]. The University of Texas, Austin.

Miller, T., Chahine, S., & Childs, R. A. (2010). Detecting differential item functioning and differential step functioning due to differences that should matter. *Practical Assessment, Research, and Evaluation, 15*(10), 1–13. https://doi.org/10.7275/dzm4-q558

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of educational measurement, 44*(3), 187–210. https://doi.org/10.1111/j.1745-3984.2007.00034.x

Penfield, R. D. (2008). Three classes of nonparametric differential step functioning effect estimators. *Applied Psychological Measurement, 32*(6), 480–501. https://doi.org/10.1177/0146621607305399

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47*(2), 129–149. https://doi.org/10.1111/j.1745-3984.2010.00105.x

Penfield, R. D. (2013). DIFAS 5.0 differential item functioning analysis system user's manual. https://soe.uncg.edu/wp-content/uploads/2015/12/DIFASManual_V5.pdf

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353–370. https://doi.org/10.1111/j.1745-3984.2003.tb01151.x

Penfield, R. D., Alvarez, K., & Lee, O. (2008). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*(1), 61–78. https://doi.org/10.1080/08957340802558367

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, 19(3), 5–15. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1111/j.1745-3992.2000.tb00033.x

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, 33(2), 215–230. https://www.jstor.org/stable/pdf/1435184.pdf

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

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, 40(2), 106–108. https://www.jstor.org/stable/pdf/2684866.pdf

Wang, W. C., & Su, Y. H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, 28(6), 450–480. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=96cc44755a12838b2cde4401a0635aaa6b075768

Wood, S. W. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small* [Unpublished doctoral thesis]. The University of Iowa, USA.

Yandı, A. (2017). *Comparison of the methods of examining measurement equivalence under different conditions in terms of statistical power ratios* [Unpublished doctoral thesis]. Ankara University, Ankara.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2), 223–233. https://doi.org/10.1080/15434300701375832

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, 2012(1), i-30. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/j.2333-8504.2012.tb02290.x

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of educational measurement*, 30(3), 233–251. https://doi.org/10.1111/j.1745-3984.1993.tb00425.x