(ISSN: 2602-4047)

Pektaş, S. & Tan, Ş. (2022). The effects of DIF on test parameters estimates, decision studies, G And Phi coefficients, *International Journal of Eurasian Education and Culture, 7(18), 1847-1863.*

DOI: http://dx.doi.org/10.35826/ijoecc.616

Article Type (Makale Türü): Research Article

THE EFFECTS OF DIF ON TEST PARAMETERS ESTIMATES, DECISION STUDIES, G AND PHI COEFFICIENTS[1]

Sami PEKTAŞ
Asst. Prof. Dr., Niğde Ömer Halisdemir University, Faculty of Education,
Department of Educational Sciences, Niğde, Turkey, pektassami@gmail.com
ORCID: 0000-0003-4753-6112

Şeref TAN
Prof. Dr., Gazi University, Gazi Education Faculty, Department of Measurement and Evaluation in Education,
Ankara, Turkey, sereftan4@yahoo.com
ORCID ID: 0000-0002-9892-3369

Received: 02.06.2022 *Accepted: 15.08.2022* *Published: 03.09.2022*

ABSTRACT

This study aims to reveal the effect of removing items with low, medium and high Differential Item Function (DIF) from numerical and verbal ability tests, which were determined by using different (DIF) detection methods, on the test parameters estimates, the G and Phi coefficients of the decision studies, and the G and Phi coefficients of the test without DIF, according to the gender of the students, their weekly pocket money amount and whether their pocket money amount is sufficient or not. In this respect, the study is a basic research. It is also a descriptive research in terms of revealing the current situation. Mantel-Haenszel and Logistic Regression Methods based on Classical Test Theory (CTT) were used as DIF detection methods, and the SIBTEST, Lord's Chi-Square and Raju's Area Measures methods were used based on Item Response Theory (IRT). In order to reveal the general abilities of the students, 45-item numerical and 45-item verbal ability tests were used in the multiple-choice test item type by the researcher. Research data were collected from 2304 sixth grade students. As a result of the research, it is seen that there is no change that creates a significant difference in measurement comparisons as a result of removing items with low DIF from the test by using different DIF detection methods. However, as a result of removing the items showing medium and high DIF from the test, it was determined that there was a significant difference in some test parameters, some reliability coefficients, G and Phi coefficients of decision studies, and G and Phi coefficients of the test without DIF. In this context, it can be said that in order to positively increase the psychometric properties of a test, such as validity and reliability, the test should be freed from items showing medium and high DIF.

Keywords: Differential item function, ability, generalizability coefficient, test parameter estimates.

[1] "The Effects of Differential Item Functioning Determination Methods on Test Parameters Estimates, Decision Studies, G And Phi Coefficients" Produced from doctoral thesis named.

INTRODUCTION

The concept of "ability", which expresses a mental power, essentially refers to a mental processing group. It is known that separate factors are effective in the mental processes of individuals and that these factors can be grouped according to common characteristics. In other words, all processes that require mental activities can be grouped. The operations in these groups require a specific and distinct mental power.

The mental power required for each group is called the basic factor or ability. Since ability differs individually and expresses a developing process, it is important to measure ability. In this direction, tests have been developed that can make the recognition of individuals possible in terms of their abilities and reveal the diversity among individuals. Ability tests are in the range of tests created to determine what individuals can do through the education they receive, and achievement tests that measure acquired skills and predict what behaviors they will exhibit at that point. In addition to ability tests that measure multiple skills, there are also ability tests that measure very special abilities (R. Atkinson, Atkinson & Hilgard, 1995).

Since abilities can be classified into two groups, general and special abilities, ability tests are also divided into two groups. While the ability tests, which are divided into, two general and special abilities, were created in a homogeneous structure, as a result of the examination of the intelligence tests used later, it was revealed that the tests not only measure all the features of the intelligence universe but also features such as language, number and reasoning, that is, they have a heterogeneous structure (Ozguven, 2007). "Many writers like Binet used the concepts of 'general level', 'general ability' and 'general intelligence' for the same purpose" (Spearman, 1927).

In this study, a test based on Spearman's two-factor theory was developed to measure students' ability in numerical and verbal domains. As predicted by this theory, the test was developed by considering a common general factor (g factor) for each test group and a specific factor (s factor) peculiar to each test during the development of the general ability test. All tests developed or administered are intended to measure a single skill or ability. This property is called “one-dimensionality”. The fact that all the items in the test come together and be one-dimensional is a prerequisite in the study of determining item bias (Kurnaz, 2006). Dorans and Holland (1993) state that the study of determining item bias begins with a statistical process. Differential item functioning (DIF) studies, in which the correct answer probability of an item at the same ability level and in different groups are examined and its functions in subgroups are compared, is the first stage of determining item bias. At this point, it is necessary to get to the bottom of the difference to determine the bias status of the items showing DIF. In subgroups, it is decided whether an item showing DIF is biased as a result of the examinations made by considering the structure or scope of the test (Atalay, Gök, Kelecioğlu & Arsan, 2012).

Many techniques have been developed for DIF detection based on Classical Test Theory (CTT) and Item Response Theory (IRT). Methods based on CTT include Logistic regression (LR), chi-square, analysis of variance and Mantel-Haenszel (MH) methods. Methods based on IRT include SIBTEST, Lord's chi-square test, Raju's Area

Measures and likelihood ratio methods. DIF detection methods are very well structured for two-category data. In tests consisting of correct/false (1-0) scored items, MH and LR methods are the most used when detecting DIF. Although these methods are not based on IRT, DIF is detected by matching individuals at the raw score level according to the MH method and at the ability level according to the LR method (Camilli & Shepard, 1994). A comparison of groups in IRT can be achieved with item characteristic curves.

In IRT, the item characteristic curve and the responses of the reference and focus groups on the same item are compared. The difference in the item characteristic curve shows that the probability of answering the item differs for people with the same ability level in different groups. Unlike CTT, comparisons are made on the basis of ability level, not group performance (Camilli & Shepard, 1994, p. 58).

The DIF result revealed in the items in the test administered to individuals, and some test items have less validity for at least one of the subgroups. In this direction, studies in the literature show that items showing DIF can be removed from the test. Narayanan & Swaminathan (1994) state that removing the items showing DIF from the test among the items in the test also increases the reliability of the test for all groups. In this direction, considering the content validity of the examined test, items with DIF should be removed from the test or revised. In this research, it was examined how the removal of the item with DIF from the test caused a change in the calculated reliability and test parameters estimates based on the CTT, Generalizability Theory and IRT. In this context, reliability calculation methods based on CTT do not clearly reveal multiple error sources in a single application.

Generalizability theory, on the other hand, is considered as a model in which multiple error sources can be handled as an extension of both CTT and analysis of variance (Güler, Uyanık & Teker, 2012, p. 3). Within the scope of the research, in addition to the estimates based on CTT and IRT, calculations of the generalizability theory were also carried out. In this direction, generalizability theory has been detailed in line with the point of view that is the subject of the study.

In generalizability theory, which allows the estimate of multiple sources of error simultaneously, there is a distinction between relative and absolute evaluations. Therefore, when calculating reliability, there is a difference between the error variances according to the relative and absolute errors (Brennan, 2001). Another feature that distinguishes G theory from classical test theory is that a decision or improvement study (K study) can be made to obtain a higher reliability coefficient with the generalizability theory (Brennan, 2001; Crocker & Algina, 1986; Güler, 2008; Rentz, 1987; Shavelson & Webb, 1991). By performing K study with generalizability theory, information is obtained for future studies to reduce error sources in measurements with scenarios created for future studies.

This study aims to remove the items with DIF from the general ability test, which are determined by using Differential Item Functioning (DIF) detection methods according to the gender of the students, the amount of weekly pocket money and whether the amount of pocket money is sufficient or not, and to find out whether there is a significant difference in the estimates of test parameters, the G and Phi coefficients of decision studies, and the G and Phi coefficients of the test without DIF. Within the scope of the research, analyzes were also carried out on the basis of DIF detection methods, and it was aimed to determine which method had the greater effect of removing the items from the test according to the number and level of DIF items. In this context, the problem situation is analyzed and explained in the following way within the framework of abilities and measuring abilities, test and item bias, and generalizability theory.

METHOD

Research Model

The aim of this study is to determine whether the removal of low, medium and high DIF items from the test among the items in the general ability tests using DIF methods differs significantly in the test parameters estimates, the G and Phi coefficients of the decision studies, and the G and Phi coefficients of the test without DIF. In this respect, the study is a basic research. It is also a descriptive research in terms of revealing the current situation. In descriptive research, it is tried to explain the relationships between the variables examined by taking into account the previous situations of the examined events (Brown, Cozby, Kee & Worden, 1999).

Study Group

Within the scope of the research, data obtained from many study groups were used for different purposes. First of all, 536 students were reached during the development phase of the general ability test, which consists of items at the level of numerical and verbal ability. In the preliminary application for the validity and reliability analysis of the test, an application was made to the sixth grade students in Keçiören and Pursaklar districts of Ankara province in the 2016-2017 academic year. During the final application of the study, since the effects of DIF detection methods on the test parameters and the reliability coefficients of the generalizability theory will be compared, the method of determining the population and sample was not used. However, a large-scale study group was formed to test DIF detection methods based on CTT and IRT. The application was made to 2304 students studying at the sixth grade level in the 2016-2017 academic year in the Keçiören and Pursaklar districts of Ankara. In studies on DIF in the literature, it is seen that the probability of correctly detecting items with DIF increases with the increase in sample size. In other words, it is stated that the probability of error of DIF detection methods decreases as the number of individuals increases (Narayanan & Swaminathan, 1994).

Table 1 shows the distribution of the sixth grade students participating in the research regarding their gender, weekly pocket money amounts and whether their pocket money amounts are sufficient or not.

Table 1. The Frequency and Percentage Distributions of the Students on Gender, Weekly Pocket Money and Whether the Amount of Pocket Money Is Sufficient

Variable	Category	DIF Classification	*f*	%
Gender	Female	Focus (1)	1146	49.7
	Male	Reference (0)	1158	50.3
Pocket Money Amount	1 (0-5 TL) Low	Reference (0)	608	26.4
	2 (6-20 TL) Medium		1118	48.5
	3 (21-120 TL) High	Focus (1)	578	25.1
Sufficiency of Pocket Money Amount	Sufficient	Reference (0)	1914	83.1
	Not sufficient	Focus (1)	390	16.9
Total			2304	100.0

In order to compare the two groups before the DIF analysis of the data collected from the students, the students are classified as focus and reference groups according to the feature that expresses the DIF source. Focus and reference groups are equated according to their numerical and verbal ability levels, and DIF methods are applied to reveal the differences between the correct answer probabilities of the items and the group differences by considering this classification (Zumbo, 1999). The reference group concept used in this study reflects the majority group, and the focus group concept reflects the minority group (Santelices & Wilson, 2012).

Data Collection Tool

50 items measuring numerical ability and 50 items measuring verbal ability were selected by the researcher within the framework of expert opinions, taking into account item discrimination and difficulty indices. The selected items were revised again, and the General Ability Test was finalized for the pre-trial application. There are two parts in the developed test, namely verbal ability and numerical ability. Each section consists of a total of 50 items. Each question has four response options, and only one of the options is coded as the correct answer.

As a result of the pilot applications in the numerical ability test, two items were removed from the 50 items before the pilot application. A 45-item numerical ability test was developed by subtracting three items from the remaining 48 items as a result of expert opinions and calculated statistics. As a result of the pilot application, it is seen that the average difficulty index of the 45-item numerical ability test is 0.51 and a medium difficulty test. Considering the upper-lower group mean item discrimination index value (0.43), it is seen that the test has the ability to distinguish students with high numerical ability from those with low numerical ability. The KR-20 reliability coefficient calculated with the data obtained from the items scored 1 and 0 on the numerical ability test was found to be 0.90. In the verbal ability test, four items were removed from 50 items before the pilot application, and after the pilot application, only the 3rd item was removed from the remaining 46 items due to misconception. As a result, a 45-item verbal ability test was designed. As a result of the pilot application, it is seen that the average difficulty index of the 45-item verbal ability test is a medium difficulty test with 0.58. Considering the upper-lower group mean item discrimination index value (0.40), it is seen that the test is at a level to distinguish students with high verbal ability from students with low verbal ability. The KR-20 reliability coefficient of the verbal ability test was calculated as 0.85.

The 45-item numerical ability test and the 45-item verbal ability test developed for sixth grade students were applied to 2304 students studying in sixth grade in schools affiliated to Keçiören and Pursaklar District National Education Directorates within the scope of the final application. During the application process, students were given 60 minutes of time in separate sessions for the 45-item numerical ability test and verbal ability tests, and student answers were collected with optical forms besides the question booklets. The reliability coefficient of KR-20 based on CTT for the numerical ability test was calculated as 0.90, and the reliability coefficient of Lord based on IRT was calculated as 0.91. The reliability coefficient of KR-20 based on CTT for the verbal ability test was calculated as 0.84, and the reliability coefficient of Lord based on IRT was calculated as 0.87.

As a result, when the item and test statistics of the 45-item tests that can be scored 1-0, which were developed and finalized to reveal the verbal abilities and numerical abilities of the sixth grade students, it is seen that the tests are valid and reliable ability tests.

Data Analysis

Analysis of Data on Pre-application

During the development of the general ability test used in the research, the items were revised by the researcher and the items with a low item discrimination index were arranged and designed for the trial application.

The draft form of the previously examined numerical and verbal ability tests was revised by the researcher, and the arrangement was made by taking expert opinions. The prepared form was applied to 536 students studying at the sixth grade level, and the TAP (Test Analysis Program Version 14.7.4) package program was used to calculate the item and test statistics based on CTT for the data. BILOG-MG (Version 3.0) (Binary Logistic Models) package program was used to calculate item and test statistics based on item response theory. 268 data were drawn randomly from the data obtained from the trial application, and a comparison was made with the other 268 data in terms of item parameters. In this context, the final application was made by ensuring the invariance of the item parameters and removing the items with great variability from the test. After these processes, the general ability test was given its final shape and data were obtained by applying it to 2304 students.

Analysis of Data on Final Application

The data were organized in the context of subgroups according to the gender of the students, their weekly pocket money and their sufficient pocket money amount, using IBM SPSS-25 and the Excel package program.

MH and LR methods based on CTT, SIBTEST based on IRT, Lord's Chi-Square and Raju's Area Measures methods were used according to gender, weekly pocket money amounts and whether pocket money amounts were sufficient or not, and the RStudio package program was used to detect DIF.

R is an open-source program distributed free of charge over the internet and can run on almost all operating systems. R libraries are developed with ready-made code and functions and provide convenience to users (Team, 2013). While the "mirt" package is loaded for SIBTEST, one of the DIF detection methods, the "difR" package is loaded in other DIF detection methods and commands are written for analysis.

Before proceeding to the DIF detection methods based on IRT, the assumptions of normality of the distribution, one-dimensionality and local independence of the preconditions of the IRT were tested. Multivariate extreme values were analyzed by calculating Mahalanobis. Tetrachoric Exploratory Factor analysis was performed with the Factor Version 10.3.01 64 Bits (Lorenzo-Seva & Ferrando, 2015) program since it is a structure scored 1-0 to test the one-dimensionality assumptions.

When these values are examined, the numerical ability test in a single factor explains 34% of the total variance. At the same time, it is seen that the eigenvalues of the first factor are 15.24 and the eigenvalues of the second factor are 2.68, among the eigenvalues examined to test the one-dimensionality assumption; in other words, since the ratio between the two factors is high and the eigenvalues of the factors coming after the second factor are lower than the second factor, it is a dominant factor in the numerical ability test. In this context, it is seen that the one-dimensionality assumption is met.

When the values of the verbal ability test for tetrachoric exploratory factor analysis are examined, the verbal ability test explains 30% of the total variance in a single factor. At the same time, among the eigenvalues examined to test the one-dimensionality assumption, the eigenvalue for the first factor was calculated as 13.28, and the eigenvalue for the second factor was calculated as 2.88. In this context, it is seen that the one-dimensionality assumption is met. The local independence assumption requires that separate answers to the items in a test be independent of each other.

Determining the one-dimensionality of a test means that the covariances for the items are zero for respondents at the same ability levels. This indicates that if the one-dimensionality assumption is met, the local independence assumption is also met (Çıtak, 2007). In this context, when the eigenvalues are analyzed as a result of the tetrachoric exploratory factor analyses based on the correlation matrix made with the data collected by using numerical and verbal ability tests, it is seen that the local independence assumption is also met due to the one-dimensionality assumption being met.

In the context of the subgroups determined regarding the gender, weekly pocket money amounts and the sufficiency of the weekly pocket money amounts of the students, the items showing low, medium and high levels of DIF were extracted separately from the numerical and verbal ability sections of the general ability test by using DIF methods based on CTT and IRT. The changes in the standard errors of the test items according to CTT and IRT were also calculated and compared with and without the DIF item in the test.

Since it is aimed to investigate whether there is a significant difference in the extraction of items showing DIF in the ability test, test parameters estimates, G and Phi coefficients related to decision studies, and G and Phi coefficients of the test without DIF in each DIF detection method, first of all, EduG 6.1e package program was used to test the decision studies and to determine the G and Phi coefficients.

The reliability comparison formula defined by Feldt (1969) was used to test the significant difference between the G and Phi coefficients related to the final test and the G and Phi coefficients related to the decision studies calculated after removing the items showing DIF and the test extended with the Spearman-Brown formula, and also to compare the KR-20 based on CTT and Lord's reliability coefficients based on IRT with the coefficients of the final test.

The average difficulty index and upper-lower group item discrimination index of the test obtained as a result of removing the items with DIF from the test using DIF detection methods based on CTT and IRT were tested with the Z statistic used to test the significance of the difference between the two ratios.

Within the scope of the research, Fisher Z statistics was used to determine whether there was a significant difference between the point biserial correlation and the discrimination coefficients of the final test based on CTT and the test consisting of the remaining items when the item with DIF was removed. Fisher Z statistic was defined in Akhun's (1984) research.

FINDINGS

Is the Change in the Estimates of the Test Parameters and the Generalizability Coefficients Significant as a result of the Removal of Low, Medium, and High DIF Items from the Test in DIF Detection Methods Based on CTT Regarding Student Variables in Numerical and Verbal Ability Test?

Summary on the effect of removal of DIF-identified items from the test using DIF detection methods based on classical test theory on test parameter estimates and generalizability coefficients in numerical and verbal ability tests Table 2 is shown.

Table 2. Summary on the Effect of Removal of DIF-Identified Items from the Test Using DIF Detection Methods Based on Classical Test Theory on Test Parameter Estimates and Generalizability Coefficients in Numerical and . Verbal Ability Tests

			A Level DIF							B Level DIF							C Level DIF						
	Variable	**Test**	**KR-20**	**p**	**q**	**$r_{nç}$**	**Lord r**	**Decision**	**S-B Test**	**KR-20**	**p**	**q**	**$r_{nç}$**	**Lord r**	**Decision**	**S-B Test**	**KR-20**	**p**	**q**	**$r_{nç}$**	**Lord r**	**Decision**	**S-B Test**
MH Method	Gender	Numerical	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	√	X	X	X	X	√	√	-	-	-	-	-	-	-
	Pocket Money	Numerical	X	X	X	X	X	X	X	√	X	X	X	X	√	√	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Sufficiency of Pocket Money	Numerical	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LR Method	Gender	Numerical	X	X	X	X	X	●	●	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	√	X	√	X	√	√	X	√	√	-	-	-	-	-	-	-
	Pocket Money	Numerical	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Verbal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Sufficiency of Pocket Money	Numerical	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-

p: average difficulty; q: upper-lower group item discrimination index; $r_{nç}$= Mean Point Biserial Correlation; Decision: Decision Study G and Phi Coefficients; S-B Test: The G and Phi coefficients of the extended test with the Spearman-Brown formula; X = No Significant Increase; √= There Is A Significant Increase; ●= There Is a Significant Decline; - Non-DIF situation.

Table 2 demonstrates that there is information summarizing whether it has increased at a level that will make a significant difference as a result of the comparison of the coefficients related to the final test with the test parameters estimates and generalizability coefficients obtained as a result of removing the items showing DIF at the A, B and C levels from the test of DIF detection methods based on classical test theory, according to students' gender, weekly pocket money amount and whether their weekly pocket money amounts are sufficient or not. It is seen that removing the items showing DIF at level A in the MH and LR methods from the test did not reveal a significant difference in the compared coefficients. However, according to the gender variable of the LR method, the removal of the item showing DIF at the A level in the numerical ability test from test caused a significant decrease in the generalizability coefficients of the decision study and the extended test. Removing the items showing DIF at the A level in the verbal ability test according to the gender variable related to the LR method caused an increase that would create a significant difference in the decision study. It is observed that removing items showing DIF at B level in the MH method and LR method from the test increased at a level that would make a significant difference in the G and Phi coefficients of the decision study and the extended test in the KR-20 reliability coefficient. In the LR method, at the same time, the removal of items showing DIF at the B level from the test resulted in an increase in the discrimination value of the upper-lower group item discrimination index and the point-biserial correlation. It is seen that the probability of detecting items with medium DIF in verbal ability tests of MH and LR methods, which are DIF detection methods based on CTT, is higher than in numerical ability tests.

Is the Change in Test Parameter Estimates and Generalizability Coefficients Significant after Removal of Low, Medium, and High DIF Items from the Test in IRT-Based DIF Detection Methods Regarding Student Variables in Numerical and Verbal Ability Tests?

Summary on the effect of removal of DIF-identified items from the test using DIF detection methods based on item response theory on test parameter estimates and generalizability coefficients in numerical and verbal ability tests Table 3 is shown.

Table 3. Summary on the Effect of Removal of DIF-Identified Items from the Test Using DIF Detection Methods Based on Item Response Theory on Test Parameter Estimates and Generalizability Coefficients in Numerical and Verbal Ability Tests

			A Level DIF							B Level DIF							C Level DIF						
	Variable	**Test**	**KR-20**	**p**	**q**	$r_{nç}$	**Lord r**	**Decision**	**S-B Test**	**KR-20**	**p**	**q**	$r_{nç}$	**Lord r**	**Decision**	**S-B Test**	**KR-20**	**p**	**q**	$r_{nç}$	**Lord r**	**Decision**	**S-B Test**
SIBTEST Method	Gender	Numerical	-	-	-	-	-	-	-	√	√	√	X	X	√	√	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	√	√	√	X	√	√	√	√	√	√	√	√	√	√
	Pocket Money	Numerical	-	-	-	-	-	-	-	√	X	X	X	X	√	√	√	X	X	X	X	√	√
		Verbal	-	-	-	-	-	-	-	√	X	X	X	√	√	√	-	-	-	-	-	-	-
	Sufficiency of Pocket Money	Numerical	-	-	-	-	-	-	-	√	X	√	X	X	√	√	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	√	X	X	X	√	√	√	-	-	-	-	-	-	-
Lord's Chi-Square Method	Gender	Numerical	X	X	X	X	X	X	X	√	√	√	X	√	√	√	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	√	X	X	X	√	√	√	√	√	√	√	√	√	√
	Pocket Money	Numerical	X	X	X	X	X	X	X	√	X	X	X	X	√	√	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	-	-	-	-	-	-	-	√	X	X	X	√	√	√
	Sufficiency of Pocket Money	Numerical	X	X	X	X	X	√	√	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	√	X	X	X	√	√	√	-	-	-	-	-	-	-
Raju's Method of Area Measures	Gender	Numerical	X	X	X	X	X	X	X	√	√	√	X	√	√	√	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	√	√	√	X	√	√	√	√	√	√	X	√	√	√
	Pocket Money	Numerical	X	X	X	X	X	X	X	√	X	X	X	X	√	√	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Sufficiency of Pocket Money	Numerical	X	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Verbal	X	X	X	X	X	X	X	√	X	X	X	√	√	√	-	-	-	-	-	-	-

p: average difficulty; q: upper-lower group item discrimination index; $r_{nç}$= Mean Point Biserial Correlation; Decision: Decision Study G and Phi Coefficients; S-B Test: The G and Phi coefficients of the extended test with the Spearman-Brown formula; X = No Significant Increase; √= There Is A Significant Increase; ●= There Is a Significant Decline; - Non-DIF status.

Table 3 shows that there is information summarizing whether it has increased at a level that will make a significant difference as a result of the comparison of the coefficients related to the final test with the test parameters estimates and generalizability coefficients obtained as a result of removing the items showing DIF at the A, B and C levels from the test of DIF detection methods based on the item response theory, according to students' gender, weekly pocket money amount and whether their weekly pocket money amounts are sufficient or not. In the numerical and verbal ability tests, it was determined that the SIBTEST method for all three variables of the students was an IRT-based method in which medium-level DIF was detected, and low-level DIF was found very rarely. In consequence of removing the items with DIF at level A, which was detected by methods based on IRT, from the test, it is seen that there is no significant difference in test parameters estimates, decision studies, G and Phi coefficients as a result of comparisons with the final test. It is seen that detecting medium and high DIF from IRT-based methods and removing it from the test causes an increase at a level that will make a significant difference as a result of the comparison of the KR-20 reliability coefficient, the G and Phi coefficients of the decision studies, the G and Phi coefficients of the extended test with the Spearman-Brown formula, with the final test values. In DIF studies of SIBTEST, Lord's Chi-square and Raju's Area Measures, it is seen that there is an item with DIF at A, B and C levels according to the gender of the students in verbal ability tests.

It is seen that there is an item with DIF at A and B levels in the methods based on IRT, according to the sufficiency of pocket money amount of the students. As a result of the comparison of the two categories, low and high, according to the amount of pocket money of the students, it was seen that there was an item with DIF at the B and C levels, which was detected by using the SIBTEST method in the numerical ability test, and that there is no significant difference, in general, as a result of comparison of the average difficulty, upper-lower item discrimination index, point-biserial correlation discrimination value, and Lord's reliability coefficients with the final test.

As a result of the SIBTEST method, it is seen that there are items with DIF at the B level in the numerical ability test according to gender and whether the amount of pocket money is sufficient, and that there is no significant difference in the context of comparing the discrimination coefficients with the point-biserial correlation. It is seen that removing items with DIF at level B based on IRT from the test in all variables does not contribute significantly to the discrimination value and the point biserial correlation.

What is the Distribution of the Number of Items with DIF Calculated in the Numerical Ability Test According to the DIF Detection Methods Based on CTT and IRT?

Number of items with DIF calculated in the numerical ability test based on dif detection methods based on CTT and IRT Table 4 is shown.

Table 4. Number of Items with DIF Calculated in the Numerical Ability Test Based on DIF Detection Methods Based on CTT and IRT

		Numerical Ability Test		
Variable	**DIF Method**	**Low**	**Medium**	**High**
Gender	MH	11	-	-
	LR	12	-	-
	SIBTEST		3	-
	Lord's Chi-Square	16	1	-
	Raju's Area Measures	9	2	-
Pocket Money Amount	MH	1	1	-
	LR	4	-	-
	SIBTEST	-	2	1
	Lord's Chi-Square	1	2	-
	Raju's Area Measures	1	1	-
Whether the Pocket Money Amount Is Sufficient	MH	1	-	-
	LR	-	-	-
	SIBTEST	-	5	-
	Lord's Chi-Square	5	-	-
	Raju's Area Measures	1	-	-

When Table 4 is examined, it is seen that DIF items with medium level are observed the most in the SIBTEST method among the DIF detection methods based on CTT and IRT in the numerical ability test according to gender. When the MH and LR methods are examined, it is seen that there is no medium or high-level DIF item in the numerical ability test according to gender. In the numerical ability test, no item with a high level of DIF was detected in any of the CTT and IRT-based DIF detection methods according to gender.

According to the pocket money amounts of the students, it is seen that there are medium and high-level DIF items in the SIBTEST method. It was determined that, only in the LR method, no items with medium or high levels of DIF were observed. According to the sufficiency of the students' pocket money amount, it is seen that there is only a medium level of DIF item in the SIBTEST method. In other methods, except SIBTEST, of the DIF detection methods based on CTT and IRT, no item with medium and high DIF was detected.

What is the Distribution of the Number of Items with DIF Calculated in the Verbal Ability Test According to the DIF Detection Methods Based on CTT and IRT?

Number of items with DIF calculated in verbal ability test according to dif detection methods based on CTT and IRT Table 5 is shown.

Table 5. Number of Items with DIF Calculated in Verbal Ability Test According to DIF Detection Methods Based on CTT and IRT

		Verbal Ability Test		
Variable	DIF Method	Low	Medium	High
Gender	MH	15	11	-
	LR	19	6	-
	SIBTEST	2	7	9
	Lord's Chi-Square	15	12	2
	Raju's Area Measures	12	8	5
Pocket Money Amount	MH	1		
	LR	-	-	-
	SIBTEST	-	1	-
	Lord's Chi-Square	1	-	1
	Raju's Area Measures	1	-	-
Whether the Pocket Money Amount Is Sufficient	MH	6	-	-
	LR	3	-	-
	SIBTEST	2	3	
	Lord's Chi-Square	7	2	-
	Raju's Area Measures	5	1	-

When Table 5 is examined, it is seen that among the DIF detection methods based on CTT and IRT in the numerical ability test, there are the most medium DIF items in the Lord's Chi-square method and the least medium DIF items in the LR method. It is seen that the method in which the most DIF items at the high level are detected according to the gender of the students is SIBTEST. When the MH and LR methods are examined, it is seen that there is no high level of DIF in the numerical ability test according to gender. According to the amount of pocket money of the students, it is seen that the DIF item was detected at a medium level in the SIBTEST method and at a high level in the Lord's chi-square method. According to the sufficiency of the pocket money amount of the students, the most medium DIF item was detected in the SIBTEST method. It is seen that no item with a high level of DIF is detected in DIF detection methods based on CTT and IRT. Based on the MH and LR methods, it is seen that there is no medium or high level of DIF items according to the sufficiency of the pocket money amount of the students.

CONCLUSION and DISCUSSION

Conclusions

In the numerical ability test, the most DIF item was detected according to the gender variables of the students. In the numerical ability test, it was concluded that the item with the most DIF at medium and high levels was detected by the SIBTEST method. In the verbal ability test, the most DIF item was detected in the gender variable, and it was concluded that while the high level of DIF item appeared in the SIBTEST method, the number of DIF items decreased in the form of Raju's Area Measures and Lord's Chi-Square methods. In the verbal ability test, the item with the most DIF at the medium level was detected in the Lord's Chi-Square method, and it was concluded that the item with the least DIF was detected in the LR Method. According to the amount of pocket money of the students, it was concluded that the method that detected the most medium and high-level DIF items in the numerical ability test was the SIBTEST method and the LR method was the least. According to the amount of pocket money of the students, it was concluded that the method that detected the most medium DIF item in the verbal ability test was the SIBTEST method, and the method that detected the high-level DIF item was the Lord's Chi-Square method. According to the sufficiency of the students' pocket money amount, it was determined that the SIBTEST method is the method that detects the item with the most DIF at the medium level, and the LR and MH methods are the least. In the context of the numerical ability test and verbal ability test, it was concluded that DIF detection methods based on IRT detected more DIF items at medium and high levels than DIF detection methods based on CTT.

It was concluded that removing the items showing medium and high DIF from the test using DIF detection methods based on CTT and IRT created a statistically significant difference in the KR-20 reliability coefficients of the test obtained and the initial final test. Likewise, when the G and Phi coefficients calculated as a result of the decision studies and the G and Phi coefficients of the tests without DIF, which were extended to the final test number by using the Spearman-Brown formula, were compared on the basis of all methods and variables, it was concluded that there was a significant difference. It was concluded that the LR method, which is one of the methods based on CTT, is a more effective method on the test parameters than the MH method.

It was concluded that Raju's Area Measures method was the method that affected the test parameters at the level that would make a statistically significant difference by helping to remove the item with medium DIF from the test in terms of gender variable, and the MH method was the least. In terms of pocket money amount variable, it was concluded that the method that affected the test parameters at the level that would make the most statistically significant difference by helping to remove the medium-level DIF item from the test was the

SIBTEST method and that Lord's chi-square method and Raju's Area Measures method are similar in influencing level to make a significant difference.

In terms of whether the weekly pocket money amount of the students is sufficient or not, it was concluded that the method that affected the test parameters at the level that would make the most statistically significant difference by helping to remove the medium-level DIF item from the test was the SIBTEST method, and that Lord's chi-square method and Raju's Area Measures method were similar in effect at a level that would create a statistically significant difference.

It was concluded that Lord's chi-square method is the DIF detection method which would increase the test parameters, the G and Phi coefficients in the decision studies and the test in which the DIF item is removed at a level that would make the most statistically significant difference by detecting the item with high level of DIF. In the study, it was concluded that the variable that increased with the least significant difference was the point biserial correlation and the discrimination coefficient, as a result of removing B and C level items from the test.

Discussion

As a result of the findings, it was determined that the items showing DIF in both numerical and verbal ability tests showed similarity in IRT-based methods. In other words, it was concluded that the items with DIF estimated based on SIBTEST, Lord's Chi-Square method and Raju's Area Measures method were similar. The fact that methods based on IRT produce consistent results with each other is similar to previous studies (Acar, 2008; Çepni, 2011; Kan et al., 2013; Karakaya & Kutlu, 2012). Grover and Ercikan (2017) mentioned in their research that researchers have so far conducted DIF studies on different groups such as ethnicity, socio-economic status and gender. In their study, Fleishman and Lawrence (2003) emphasized that it is important to consider the potential effect of item bias in comparing the health status of groups with different socio-economic levels in health studies. In this context, the research revealed that there are items showing DIF in favor of the upper group, especially in some items related to socio-economic status, at the stage of taking the amount of pocket money that students spend weekly as a source of DIF. In this study, it was concluded that the MH method detected more DIF items than the LR method, since the numerical and verbal ability tests did not have a complex structure and were one-dimensional. In Huang's (1998) study, it was determined that the MH method gave more reliable results than the LR method in detecting the items with DIF.

Since the items showing DIF in the literature pose a threat to the validity and reliability of the test, it is recommended that these items be removed from the test considering the content validity (Clauser & Mazor, 1998; Zumbo, 2007). It was concluded that removing items showing DIF at level A from the test with DIF detection methods based on CTT had no effect on test statistics in general, but removing items showing DIF at level B from the test increased the reliability other than Lord's reliability coefficient. Uzun and Gelbal (2017) also state that the high number of DIF items in the test makes the reliability of the test questionable. In the study of Zumbo (2000), it was revealed that the alpha coefficient and item-total correlation coefficients would increase slightly as a result of removing the items with DIF from the test, and the standard error value of the conditional measurement would decrease as expected.

In their study, Beinicke et al. (2014) mentioned that DIF is an important method to determine unbiasedness in tests and that it is necessary to remove items with DIF. In this context, it can be said that in order to positively increase the psychometric properties of a test, such as validity and reliability, the test should be purged of items showing medium and high DIF, taking into account the content validity.

RECOMMENDATIONS BASED ON THE STUDY FINDINGS

Since the results obtained from different DIF detection methods vary, it is recommended that researchers use more than one method simultaneously in the DIF detection process. In order to increase the test reliability and reduce the standard error rates of the measurement, it is recommended to remove the items showing DIF at B and C levels from the test by using DIF detection methods, and paying attention to the content validity. It is recommended that DIF analyzes be done in the process of detecting the validity and reliability of the tests applied to make serious decisions about the individual.

ETHICAL TEXT

In this article, the journal writing rules, publication principles, research and publication ethics, and journal ethical rules were followed. Responsibility for any violations that may arise regarding the article belongs to the author. The ethics committee approval for the article was granted by the Ankara Governorship National Education Directorate with the decision dated 06.12.2016 and numbered 14588481-605.99-E.13737481 by the application permission of the Gazi University Institute of Educational Sciences.

Author(s) Contribution Rate: In this study, the contribution rate of the first author is 50% and the contribution rate of the second author is 50%.

REFERENCES

Acar, T. (2008). *Determination of a differential item functioning (dif) procedure using the hierarchical generalized linear model: A comparison study with logistic regression and likelihood ratio procedure.* [Doctoral Thesis]. https://tez.yok.gov.tr accessed from.

Akhun, İ. (1984). İki korelasyon katsayısı arasındaki farkın manidarlığının test edilmesi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *17*(1), 1-7. https://doi.org/10.1501/Egifak_0000001034.

Atalay, K., Gök, B., Kelecioğlu, H., & Arsan, N. (2012). Comparing different diffential item functioning methods: A simulation study. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *43*(43), 270-281

Atkinson, R., Atkinson, R., & Hilgard, E. (1995). *Psikolojiye giriş II.* Sosyal.

Beinicke, A., Pässler, K., & Hell, B. (2014). Does gender-specific differential item functioning affect the structure in vocational interest inventories?. *International Journal for Educational and Vocational Guidance*, *14*(2), 181-198. https://doi.org/10.1007/s10775-013-9254-y

Brenan, R. L. (2001). *Generalizability theory*. Springer- Verlog.

Brown, W.K., Cozby, C.P., Kee, D.W., & Worden, P.E. (1999). *Research methods in human development* (2.b.). Mayfield.

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

Çepni, Z. (2011). *Differential item functioning analysis using SIBTEST, Mantel Haenszel, logistic regression and item Response Theory Methods.* [Doctoral Thesis]. https://tez.yok.gov.tr accessed from.

Çıtak, G.G. (2007). A comparison of differential scoring methods for multiple-choice tests in terms of classical test and itemresponse theories [Doctoral Thesis]. https://tez.yok.gov.tr accessed from.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Mason.

Dorans, N.J. & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. Differantial item functioning. P.W. Holland & H. Wainer (Ed.), *Differential item functioning* içinde (ss. 35-66). Lawrence Erlbaum.

Feldt, L.S. (1969). A test of the hypothesis that cronbach's alpha or Kuder-Richardson coefficent twenty is the same for two tests. *Psychometrika*, *34*(3), 363-373. https://doi.org/10.1007/BF02289364

Fleishman, J.A., & Lawrence, W.F. (2003). Demographic variation in SF-12 scores: true differences or differential item functioning? *Medical Care*, *41*(7), 75-86. DOI: 10.1097/01.MLR.0000076052.42628.CF

Grover, R.K. & Ercikan, K. (2017): For which boys and which girls are reading assessment ıtems biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education, 30*(3), 178-195. https://doi.org/10.1080/08957347.2017.1316276

Güler, N. (2008). A research on classical test theory generalizaibility theory and rasch model [Doctoral Thesis]. https://tez.yok.gov.tr accessed from.

Güler, N., Uyanık, K.G. & Teker, T.G., (2012). *Genellenebilirlik kuramı.* Pegem.

Huang, C.Y. (1998). *Factors influencing the reliability of DIF detection methods.* Annual Meeting of the American Educational Research Association'da sunulmuş bildiri,

Kan, A., Sünbül, Ö., & Ömür, S. (2013). 6.-8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *9*(2), 207-222

Karakaya, İ. & Kutlu, Ö. (2012). Seviye Belirleme Sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi. *Eğitim ve Bilim Dergisi, 37*(165), 348-362

Kurnaz, F.B. (2006). *Assessing item biased of peabody picture vocabulary test* [Master Thesis]. https://tez.yok.gov.tr accessed from.

Lorenzo-Seva, U. & Ferrando, P.J. (2015). *Factor version 10.3.01*. Tarragona

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328. https://doi.org/10.1177/014662169401800403.

Özgüven, İ. E. (2007). *Psikolojik testler.* Nobel.

Rentz, J. O. (1987) Generalizability theory: A comprehensive method for assessing and improving the dependability of marketing measures. *Journal Of Marketing Research*, *24*(1), 19-28. https://doi.org/10.2307/3151750

Santelices, M.V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement, 72*(1), 5-36. https://doi.org/10.1177/0013164411412943

Shavelson, R.J., & Webb, M.N. (1991). *Generalizability theory a prime*. Sage.

Spearman, C. (1927). *The abilities of man.* Melbourne

Team, R. (2013). R development core team. *RA Lang Environ Stat Comput, 55,* 275-286.

Uzun, N.B., & Gelbal, S. (2017). An investigation of item bias in PISA science test in terms of the language and culture. *Kastamonu Education Journal, 25*(6), 2427-2446.

Zumbo, B.D. (2000). *The effect of DIF and impact on classical test statistics: undetected DIF and impact, and the reliability and interpretability of scores from a language proficiency test.* Annual conference of the National Council on Measurement in Education (NCME)'de sunulmuş bildiri, April, New Orleans, LA.

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