SUJE
ISSN: 2146-7455

Received : 20.02.2023
Accepted : 12.07.2023
Available Online : 30.08.2023

SAKARYA UNIVERSITY

The Effect of Option Differences on Psychometric Properties of Items in Likert-Type Scales

Nuri DOĞAN[a] Meltem YURTÇU[b] Ceylan GÜNDEĞER[c]

a: 0000-0001-6274-2016 Hacettepe University, Turkiye nuridogan2004@gmail.com
b: 0000-0003-3303-5093 Inonu University, Turkiye meltem.yurtcu@gmail.com
c: 0000-0003-3572-1708 Aksaray University, Turkiye cgundeger@gmail.com

Abstract

Likert-type scales are often used in education and psychology. In Likert-type scales, response options/categories, like items, are expected not to direct individuals' responses. Although the researchers themselves make decision on how to arrange categories during scale development, it is possible that different categories reveal different response behaviors. In the literature, it has been observed that differentiations in the number of categories of forms are studied more, yet there are a limited number of studies investigating the middle category in the forms with different labels. Furthermore, it has also been observed that there are limited number of empirical studies conducted based on polytomous Item Response Theory. This study, which was conducted to close this gap in the literature, was carried out with 377 students. The options of the attitude scale were denominated with different labels, and thus four different forms were generated. Only the middle category names were changed in the first three forms, and in the fourth form, the categories were graded. The data obtained from the forms were analyzed using the Graded Response Model and the Generalized Partial Credit Model depending on Item Response Theory. After the examination of reliability of the forms, the parameters in these forms, and the relationships between the parameters according to both models, inferences were made as to how the differences of the middle category in the forms had an effect on the perceptions of individuals.

Keywords

Likert type scales, scale categories, perception, item response theory models

Ethics Committee Approval: Ethics committee permission for this study was obtained from İnönü University Scientific Research and Ethics Committee with the decision dated 02.07.2021 and numbered 2021/13-19.

Suggested Citation: Doğan, N., Yurtçu, M., & Gündeğer, C. (2023). The effect of option differences on psychometric properties of items in likert-type scales. *Sakarya University Journal of Education, 13*(2), 207-237. doi: https://doi.org/10.19126/suje.1253876

INTRODUCTION

Every individual has different characteristics, and perceives events differently. As living conditions and past experiences of people differ, their perceptions may differ too (Erkuş, 2012). Perception process is deemed the process of knowledge acquisition for individuals (Rajamanickam, 2007). Organizing the situations that individuals have encountered in different environments and conditions, organizing the information they have acquired affectively and becoming aware of this information is called "perception" (Dunkel, 2015; Gibson, et al., 1996; Pomerantz, 2003; Qiong, 2017). To reveal perceptions, Likert-type scales are often preferred (Wakita, Ueshima & Noguchi, 2012). These scales provide advantage by presenting numerical data about a structure (Annett, 2002), and assume a linear relationship between the response probability and the psychometric characteristic underlying the characteristic to be measured (Hulin, Dragow, & Parsons, 1983). Besides that, if the psychological distance between the categories of the characteristic to be measured is equal, precise measurements can be obtained for the structure to be measured (Wakita, Ueshima, & Noguchi, 2012).

Multiple scoring is made on polytomous categories in Likert-type scales. When a 5-point Likert-type scale is used, the answers are scored between Strongly agree (5) and Strongly disagree (1). In case that the scale is unidimensional and multi-scored, among the unidimensional Item Response Theory (IRT) models, those developed for polytomous items are utilized. It is possible to come across studies in education and social sciences in which the difference between categories of the Likert-type scales' items, which are mostly preferred due to their psychometric characteristics, is examined with polytomous IRT models (Carle et al., 2009; Cordier et al., 2019; OECD, 2021). The commonly used polytomous IRT models are Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), Graded Response Model (GRM), and Generalized Graded Response Model (GGRM). These models differ by the research aim, the parameters they use, and the fixed value of the parameters (Dai, et. al., 2021) The present study examines the forms of a 5-point Likert-type scale with different answer categories based on GRM and GPCM. The GRM was developed by Samejima (1969), and is used when item categories are sequential. Because GRM was developed for multi-scored items as an extension of the Two Parameter Logistics Model (2PLM), the model calculates a discrimination "a" parameter for each item and four threshold values that are "b1", "b2", "b3", and "b4". GPCM shows a similar structure to GRM (Bartolucci, et al. 2015; Sung & Kang, 2006). GPCM, deemed as an extended version of PCM that was developed by Masters (1982), allows parameter "a" to be different for each item (Muraki, 1992). Although GRM and GPCM are among the polytomous IRT models, and have the same number of parameters, they model the answers given to the items in different ways (Embretson & Reise, 2000).

It is also important to denominate the categories in the scales as the items in the Likert-type scales are important in terms of the measured characteristic. Categories may differ by the characteristic measured, the roots of the items used when measuring the characteristic, and the choice of the person who developed the scale. Different denominations in these categories may cause directing individuals' perceptions of the desired characteristic to be measured to extreme values, or intermediate values, or limitation of expressing perceptions (Albaum, 1997). Even the differentiation in the denomination of the middle category only can be the reason for this differentiation. While these categories are preferred, besides providing face validity, the categories being in a format that will not direct the perceptions of individuals in revealing the characteristic related to the subject to be examined while selecting these categories is important to obtain unbiased outcomes. Therefore, it is necessary that researchers exercise due diligence to determine categories in the process of scale development.

This study aims to reveal the effect of the categories in the 5-point Likert-type scales on the exhibited affective behavior, and so, on the perceptions of individuals. To that end, the answers with the same items/expressions to four forms with different scale categories were evaluated based on polytomous IRT models. In the literature, there are many studies conducted to see the effects of different expressions of scale categories (Blumberg, DeSoto, & Kuethe, 1966; Dixon, Bobo, & Stevick, 1984; Finn, 1972; Huang, 2016; Jacko & Huck, 1974; Kottner, et. al., 2011; Krosnick & Berent, 1993; Lange, et. al., 2017; Newstead & Arnold, 1989; Wetzel, et. al., 2016; Wyatt & Meyers, 1987). As, the rarity of the empirical research in which these categories are examined depending on IRT, this study can contribute to the literature. The study's problem statement is "How does the difference in the perceptions of individuals affect the item parameters according to the four forms of the same scale created with different category labels?" Based on this problem statement, answers are sought to the following sub-problems:

1. Which IRT model is more suitable for Form 1-2-3-4?
2. What are the item parameters of Form 1-2-3-4 estimated based on GPCM?
3. What are the reliability coefficients of Form 1-2-3-4 estimated based on GPCM?
4. What are the item parameters of Form 1-2-3-4 estimated based on GRM?
5. What are the reliability coefficients of Form 1-2-3-4 estimated based on GRM?
6. What is the relationship between the item parameters estimated based no GPCM and GRM for the Form 1-2-3-4?
7. What is the difference between the item parameters estimated based on GPCM for the Form 1-2-3-4?
8. What is the difference between the item parameters estimated based on GRM for the Form 1-2-3-4?

METHOD

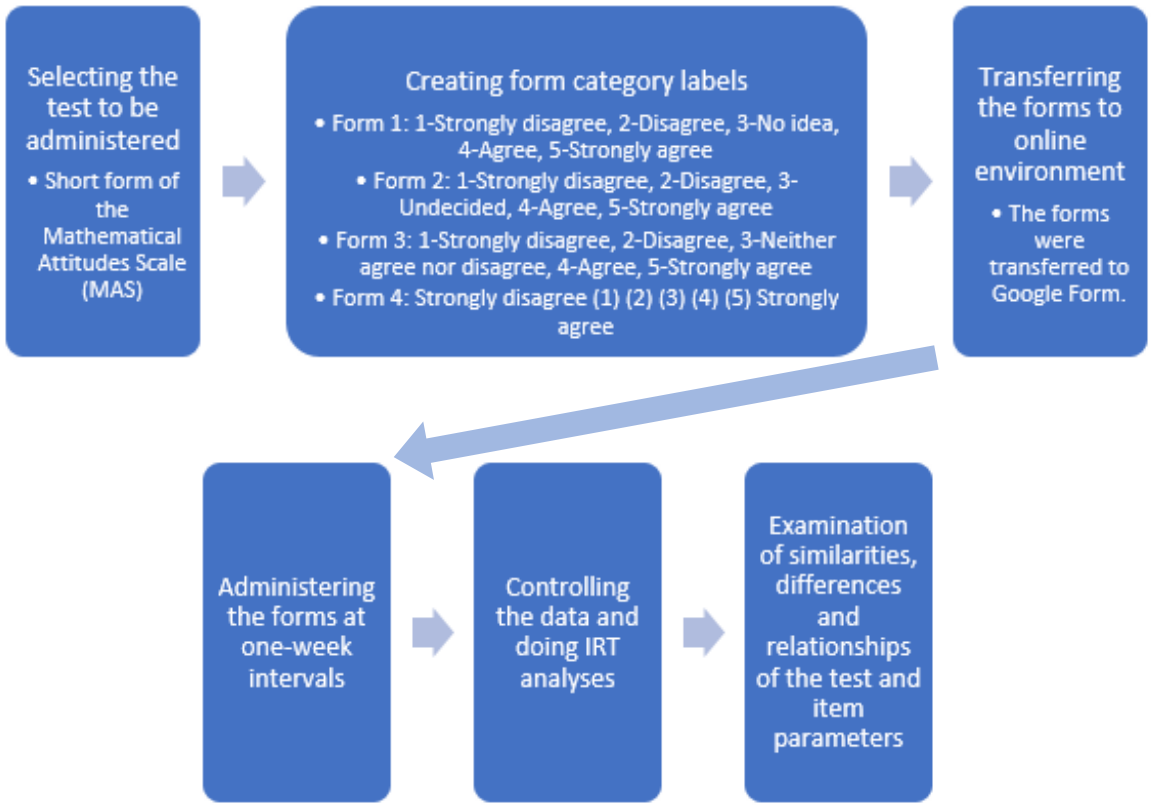
This section discusses the research design, study group, data collection tool as well as the forms created by differentiating the options, and data analysis methods in detail.

Research Design

This study was conducted based on the steps given in Figure 1. Figure 1 shows that the measurement tool was selected in the first step. Then, the category labels were changed, Form 1, 2, 3 and 4 were created, and transferred to the web environment. The forms were administered at one-week intervals, the data sets were checked, and analyses were made. This research is a descriptive study that comparatively analyzed model-data fit based on different models and different forms.

Figure 1

Research Design



Study Group

The study group consists of 377 undergraduate students from several universities of Turkey in 2021. Demographic information regarding the study group is given in Table 1. According to Table 1, almost half of the study group consists of Inonu University students (49.1%), and most of them are enrolled in the Faculty of Education (91.8%). The majority of the students who filled out the forms were at the 3rd and 4th grades, and were female students (72.1%).

Table 1

Demographic Information of the Study Group (N=377)

	n	%
Gender		
Female	272	72.1
Male	105	27.9

University		
Aksaray University	132	35
Hacettepe University	54	14.3
Inonu University	185	49.1
Other	6	1.6
Faculty		
Education Faculty	346	91.8
Sports Faculty	25	6.6
Other	6	1.6
Year		
1st Year	57	15.1
2nd Year	138	36.6
3rd Year	161	47.2
4th Year	21	5.6

Data Collection

Before the collection of data, the necessary permits were obtained from the Inonu University Scientific Research and Ethics Committee (Protocol No: 13-19, Date: 02/07/2021). Due to the COVID-19 pandemic, the forms were delivered online platform. Each form was opened to students at one-week intervals only within the administration period (four days). To minimize the error caused by the sequence effect while filling out the forms, the order of administering the forms was kept constant considering that it would be very difficult to collect data online with differing order of precedence of the forms. Thus, data obtained from 633 students filling out Form 1, 571 students filling out Form 2, 576 students filling out Form 3, and 581 students filling out Form 4 were examined. After the missing values and repetitive data in the forms were removed from the data set, analyzes were made on the data of 377 students who answered all the forms. The adequacy of the sample size for parameter estimations according to the models may differ according to the number of items in the scale and the number of categories of the items (Huang, 2016; Jin & Wang, 2014; Wetzel, et al.,2016). In the literature, for tests with 5 categories and no longer length, there are studies supporting that the sufficient sample size for the polytomous IRT model is between 300 and 500 (Kieftenbeld & Natesan, 2012), or at least 250 (Finch & French, 2019). Therefore, the number of samples used was considered sufficient to make comparisons in the model.

Data Collection Tools

In this study, the short form of the Mathematical Attitudes Scale (MAS) developed by Baykul (1990) was used as the data collection tool. The reliability coefficient of the scale with 15 positive and 15 negative items in its original form was found to be 0.96 (as cited in Nartgün, 2002, p.47). Considering that the items in the scale represent the same structure, 15 items were reviewed at the first step. Two items that were related to the mathematics of the modeled scale, and that could contribute to the students' ability to use mathematics were added to these items, and the two items stated for the lesson only were removed from the scale form. The finalized scale was organized in four different forms including four different option categories with the scale items remaining the same. For the first three forms, denomination given only for the middle values differed, while the categories of the fourth form consisted of numerical values. These categories included in the four forms are as follows:

- *Form 1:* 1-Strongly disagree, 2-Disagree, 3-No idea, 4-Agree, 5-Strongly agree
- *Form 2:* 1-Strongly disagree, 2-Disagree, 3-Undecided, 4-Agree, 5-Strongly agree
- *Form 3:* 1-Strongly disagree, 2-Disagree, 3-Neither agree nor disagree, 4-Agree, 5-Strongly agree
- *Form 4:* Strongly disagree (1) (2) (3) (4) (5) Strongly agree

Data Analysis

In this study, the analyses were conducted with unidimensional and polytomous IRT models. Therefore, the assumptions of unidimensionality and local independence of IRT of the data sets were examined first. Unidimensionality refers to a single latent trait that items measure and underlies the response performance of individuals. What is meant by unidimensionality is that there is a dominant dimension measured by the items (Hambleton & Swaminathan, 1985). In the examination of unidimensionality, the literature has offered many experimental and statistical methods. This study used principal components factor analysis using polychoric correlation calculated on polytomous item-response patterns to examine the unidimensionality assumption. For this reason, the assumptions of the principal components factor analysis, which has the purpose of explaining the variances of the measured variables rather than explaining the fundamental nature of the correlations between the measured variables, were checked on the data sets.

Table 2

KMO Values and The Bartlett Test Results

Forms	KMO	Bartlett's Chi-square	p
Form 1	0.969	42.98	.00
Form 2	0.968	38.46	.00
Form 3	0.966	37.42	.00
Form 4	0.963	22.65	.04

In the factor analysis, Comrey and Lee (1992) stated that 200 observations were suitable, 300 were good-satisfactory, and 500 were quite sufficient (as cited in Tabachnick & Fidel, 2007). The calculated KMO values and Bartlett test results for 377 people who answered the four forms are given in Table 2. Considering the KMO and Bartlett values in Table 2, the sample size is sufficient, the structure provides the assumption of multivariate normality according to Mardia's multivariate skewness and kurtosis coefficients, and it is factorable. The outliers in the data sets were examined with using the z scores and Mahalanobis, and 40 individuals in total were excluded from the data set. As obtained from the factor analysis conducted on 337 students, all forms pointed to a unidimensional structure, and the first factor in each form accounted for 76.9%, 78.7%, 79.7%, and 76.8% of the total variance, respectively. The item factor loads were between 0.73 and 0.94 for the Form 1, between 0.71 and 0.95 for the Form 2, between 0.74 and 0.95 for the Form 3, and between 0.71 and 0.95 for the Form 4, which were quite high. The Cronbach's Alpha coefficient for all forms was calculated as 0.98. The McDonald's Omega coefficients were obtained as 0.972 for Form 1, 0.975 for Form 2, 0.977 for Form 3, and 0.974 for Form 4, respectively. Based on these findings, the data on the forms supported unidimensional structure, and had high level of reliability.

Local independence is that student responses to different items are statistically independent of each other. Therefore, two items at a fixed competence level should be independent of each other (Hambleton & Swaminathan, 1985). However, it is possible that the assumption of local independence cannot be achieved in tests containing testlets, in personality tests with items representing each other, in performance assessments, and speed tests (Embretson & Reise, 2000). In such cases, because a second dimension may get involved, the assumption of unidimensionality is violated along with local independence. Because this study examines different forms of an attitude scale with different options, there is no threat to the local independence. Besides, according to McDonald, any meaningful clarity of unidimensionality is based on the principle of local independence. According to McDonald, the test is unidimensional if the covariance between items equals zero for students of similar theta levels (as cited in Hambleton & Swaminathan, 1985, p.25). In other words, if a test shows unidimensionality with a significant clarity, the items in this test also have local independence. As a result of the factor analysis, the unidimensionality of the scale can also be interpreted as the local independence of the scale items. Prior to the findings of the sub-problems, the percentage values of the options/categories in four different forms were examined, and presented in Table 3.

Table 3

The Percentages of the Categories in Forms

	Form 1					Form 2					Form 3					Form 4				
Items	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
Item 1	14	28	15	32	11	12	30	17	29	11	14	26	20	27	13	19	21	18	23	18
Item 2	6	19	16	42	16	6	17	19	41	18	6	16	22	41	15	12	16	17	31	24
Item 3	10	23	11	37	20	10	22	16	36	16	12	22	16	36	15	14	20	19	27	20

Item 4	19	31	11	27	12	16	31	16	25	11	17	31	19	23	10	24	21	18	22	15
Item 5	10	22	10	38	20	10	17	13	40	20	10	17	9	38	25	11	12	15	29	33
Item 6	21	34	11	19	15	21	30	15	19	15	23	26	18	18	15	25	23	16	16	20
Item 7	10	15	8	30	37	9	14	9	33	34	10	15	10	31	34	13	12	8	24	43
Item 8	16	21	11	27	24	14	22	11	29	23	16	21	15	27	21	19	20	14	19	28
Item 9	15	31	14	25	15	15	28	18	25	13	16	28	20	24	11	21	22	19	23	15
Item 10	12	40	12	26	9	11	36	18	25	9	13	30	23	25	9	14	29	22	23	13
Item 11	5	18	14	43	20	6	18	18	40	19	7	15	20	38	20	11	13	19	27	31
Item 12	8	18	7	40	27	8	18	10	39	25	8	16	12	39	24	12	15	12	23	38
Item 13	7	25	18	36	16	8	21	19	38	14	10	21	18	38	14	12	18	20	28	22
Item 14	12	20	8	38	21	10	23	11	34	21	11	21	13	34	21	18	13	13	28	28

Table 3 shows that students' answers have intensity in different categories and different forms. Therefore, differences occur in the response patterns of individuals in four different forms. Predictions of item parameters were first performed to test how the response differences in the forms affected the item parameters. The item parameters of the data scored with 5-point Likert-type scale generally consist of sequential response categories. The GPCM and GRM were used together in the present study to estimate item parameters of the sequential response parameters. Because these both models use item discrimination and category threshold parameters (Ostini & Nering, 2006), it is difficult to determine which model should be preferred (Dai et al., 2021). The parameters were compared based on the forms after the item parameters were estimated based on both models.

The values of -2Likelihood difference (-2LL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were examined to determine which model was more suitable for the data obtained from the forms in the solution of the first sub-problem. Lower values indicated a more appropriate model. For the second, third, fourth, and fifth sub-problems, item analyses were performed depending on GRM and GPCM, and the marginal reliability coefficients of the forms were calculated. In the sixth sub-problem, the relationship between the estimated item parameters based on two models was determined with the help of the Spearman correlation coefficient. In the seventh and eighth sub-problems, first, the Friedman test was administered to compare the estimated item parameters, and then the Wilcoxon test was administered to do pairwise comparisons. Data analysis was conducted in R software (R Development Core Team, 2013). The "*mirt*" (Chalmers, 2012) to estimate for item parameters, "mvn" (Korkmaz, Goksuluk, & Zararsiz, 2014) for multivariate normality analysis, "*TAM*" (Robitzsch, Kiefer, & Wu, 2021) for model-data fit, "*FSA*" (Ogle, Wheeler, & Dinno, 2021) for comparison between models, "*EFAtools*" (Steiner & Grieder, 2020) to test the factor structure, and "*agricolae*" (Mendiburu, 2021) packages were also benefited.

Ethical Principles

Ethics committee permission for this study was obtained from İnönü University Scientific Research and Ethics Committee with the decision dated 02.07.2021 and numbered 2021/13-19.

FINDINGS

Findings as to which IRT model is more suitable for Form 1-2-3-4

In the first sub-problem of the research, the -2LL, AIC, and BIC values were evaluated to find out fit of data with GPCM and GRM. The values obtained from each form with two separate models are given in Table 4.

Table 4

The Model Data Fit Results of the Forms

Model/Form	Likelihood	-2LL	AIC	BIC
GRM_{Form1}	-5424.388	-166.704	10988.776	11264.033
$GPCM_{Form1}$	-5507.740		11155.481	11430.738
GRM_{Form2}	-5355.097	-192.152	10850.194	11125.451
$GPCM_{Form2}$	-5451.173		11042.347	11317.604
GRM_{Form3}	-5513.108	42.04	11166.217	11441.474
$GPCM_{Form3}$	-5492.088		11124.177	11399.434
GRM_{Form4}	-5932.855	59.278	12005.710	12280.967
$GPCM_{Form4}$	-5903.216		11946.433	12221.690

When the Table 4 is examined, considering the smaller AIC and BIC values, it is seen that the model fit of GRM for Form 1 and Form 2 was better, yet Form 3 and Form 4 were found to be more compatible with GPCM. According to -2LL values, the model data fit supported this finding. Therefore, the data obtained from Form 1 with the statement "*I have no idea*" and Form 2 with the statement "*Undecided*" are more compatible with GRM, whereas the data obtained from Form 3 with the statement "*Neither agree nor disagree*" and the numerically scored Form 4 are more compatible with GPCM.

Findings as to the item parameters of Form 1-2-3-4 estimated based on GPCM

To find an answer to the second sub-problem, the discrimation (*a*) and threshold parameters (*b1, b2, b3, b4*) of the items were estimated based on GPCM for all forms, and the results were given in Table 5.

Table 5 shows that the highest standard deviation value for the discrimination parameter (a) is in Form 4, and the lowest standard deviation is in Form 1 when considering the item parameters estimated based on GPCM. a parameter values range from 0.80 to 3.92 in Form 1, from 0.76 to 4.71 in Form 2, from 0.87 to 5.24 in Form 3, and from 0.63 to 5.58 in Form 4. The discrimination values range from -2 to +2 in general, and can have infinite value. Considering the a parameters estimated from the forms, item 3 has the highest and item 8 has the lowest discrimination. The measures of central tendency (mean and median) of the discrimination index are the lowest in Form 4 and the highest in Form 2.

Table 5

The Item Parameters Estimated with GPCM

Form 1						Form 2				
	a	*b1*	*b2*	*b3*	*b4*	*a*	*b1*	*b2*	*b3*	*b4*
Item 1	3.27	-1.15	-0.10	0.05	1.34	3.68	-1.20	-0.11	0.18	1.30
Item 2	1.59	-1.92	-0.46	-0.59	1.29	1.68	-1.89	-0.65	-0.50	1.17
Item 3	3.92	-1.30	-0.33	-0.32	0.88	4.71	-1.29	-0.43	-0.10	1.03
Item 4	3.11	-0.94	0.20	0.09	1.27	4.42	-0.97	-0.01	0.31	1.28
Item 5	2.01	-1.45	-0.11	-0.64	1.02	1.48	-1.44	-0.34	-0.75	1.11
Item 6	2.54	-0.89	0.41	0.21	1.09	3.33	-0.80	0.16	0.39	1.07
Item 7	2.63	-1.35	-0.43	-0.77	0.31	3.27	-1.33	-0.56	-0.64	0.42
Item 8	0.80	-1.09	0.53	-1.01	0.73	0.76	-1.33	0.58	-1.11	0.85
Item 9	1.99	-1.22	0.18	0.01	1.16	2.68	-1.08	-0.06	0.22	1.22
Item 10	2.70	-1.33	0.30	0.16	1.50	3.39	-1.27	0.03	0.33	1.42
Item 11	1.62	-2.06	-0.50	-0.78	1.09	1.89	-1.89	-0.63	-0.47	1.09
Item 12	2.23	-1.60	-0.25	-0.91	0.71	2.10	-1.60	-0.34	-0.78	0.76
Item 13	1.71	-1.89	-0.27	-0.32	1.25	1.77	-1.69	-0.42	-0.30	1.39
Item 14	3.17	-1.23	-0.27	-0.52	0.86	3.35	-1.31	-0.28	-0.30	0.84
Mean	2.37	-1.38	-.07	-.38	1.03	2.75	-1.36	-.21	-.25	1.06
Median	2.38	-1.31	-.18	-.42	1.09	2.97	-1.32	-.31	-.30	1.10

St. Dev.	.83	.36	.34	.42	.31	1.16	.31	.34	.48	.27
	Form 3					Form 4				
	a	*b1*	*b2*	*b3*	*b4*	*a*	*b1*	*b2*	*b3*	*b4*
Item 1	4.28	-1.09	-0.22	0.24	1.17	3.19	-0.85	-0.21	0.15	0.93
Item 2	1.69	-1.79	-0.75	-0.36	1.33	1.78	-1.14	-0.55	-0.32	0.76
Item 3	5.24	-1.14	-0.37	-0.04	1.06	5.58	-1.05	-0.41	0.06	0.84
Item 4	3.92	-0.94	0.02	0.40	1.33	3.22	-0.70	-0.10	0.29	1.11
Item 5	1.74	-1.40	-0.19	-0.88	0.81	1.50	-1.15	-0.71	-0.56	0.37
Item 6	2.89	-0.74	0.05	0.44	1.03	2.16	-0.64	0.04	0.33	0.76
Item 7	2.45	-1.28	-0.48	-0.62	0.43	1.82	-1.05	-0.44	-0.80	0.02
Item 8	0.87	-1.09	0.17	-0.56	0.91	0.63	-0.67	0.27	-0.31	-0.08
Item 9	2.99	-1.02	-0.06	0.32	1.29	2.89	-0.78	-0.15	0.24	1.08
Item 10	2.84	-1.20	-0.11	0.37	1.47	2.80	-1.16	-0.14	0.32	1.20
Item 11	1.73	-1.65	-0.75	-0.41	1.04	1.39	-1.20	-0.82	-0.33	0.39
Item 12	2.39	-1.47	-0.49	-0.61	0.79	1.55	-1.15	-0.45	-0.55	0.12
Item 13	1.97	-1.47	-0.37	-0.26	1.32	1.87	-1.23	-0.49	-0.13	0.82
Item 14	3.31	-1.28	-0.36	-0.27	0.84	1.74	-0.72	-0.47	-0.46	0.58
Mean	2.73	-1.25	-.27	-.16	1.05	2.29	-.96	-.33	-.14	.63
Median	2.64	-1.24	-.29	-.26	1.05	1.84	-1.05	-.42	-.22	.76
St. Dev.	1.17	.28	.28	.44	.28	1.20	.22	.29	.37	.41

Considering the threshold parameters in Table 5, the highest standard deviation for the threshold parameter “b1” is in Form 1, and the lowest one is in Form 4. The highest central tendency values are observed in Form 4, while the lowest values are in Form 1 for the mean and in Form 2 for the median. The estimated b1 values range from -2.06 to -0.89 in Form 1, from -1.89 to -0.80 in Form 2, from -1.79 to -0.74 in Form 3, and from -1.23 to -0.64 in Form 4. The highest standard deviation for the threshold parameter “b2” was obtained in Form 1 and Form 2, and the lowest was obtained in Form 3. Of the tendency measures for this parameter, the highest value was obtained from Form 1, and the lowest

was obtained from Form 4. The estimated "b2" range from -0.50 to 0.53 in Form 1, from -0.65 to 0.58 in Form 2, from -0.75 to 0.17 in Form 3, and from -0.82 to 0.27 in Form 4.

Table 5 shows that the lowest standard deviation for the threshold parameter "b3" was obtained from Form 4. The highest standard deviation value was calculated in Form 2. Considering the central tendency measures of the "b3" parameter, the highest values were found in Form 4, and the lowest were in Form 1. The values of this parameter range from -1.01 to 0.21 in Form 1, from -1.11 to 0.39 in Form 2, from -0.88 to 0.44 in Form 3, and from -0.80 to 0.33 in Form 4. The lowest standard deviation was obtained from Form 2, and the highest was obtained from Form 4 for the threshold parameter "b4". The highest tendency measures were found in Form 2, and the lowest ones were in Form 4. Parameters "b4" range from 0.31 to 1.50 in Form 1, from 0.42 to 1.42 in Form 2, from 0.43 to 1.47 in Form 3, and from -0.08 to 1.2 in Form 4. Based on this finding, Form 1 with the statement "*I have no idea*" in the middle value and Form 2 with the statement "*Undecided*" showed similar values in similar parameters. Form 4 had the highest value in parameter "b1" and the lowest value in parameter "b4".

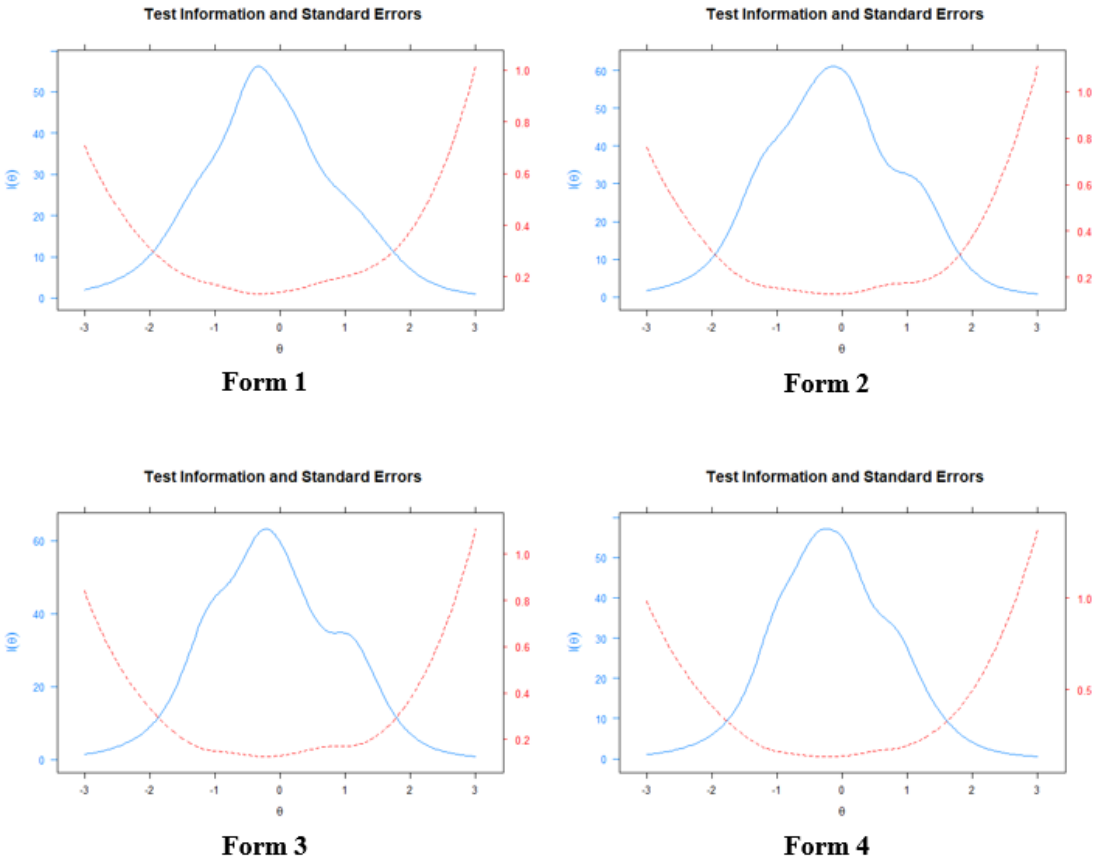
Findings as to the reliability coefficients of Form 1-2-3-4 estimated based on GPCM

In the third sub-problem, test information functions, which reflected the overall of the parameters estimated based on GPCM, and were equal to the sum of the item information functions at the relevant competence level, were obtained. Test information functions provides information about reliability on IRT. Thus, high level of test information showed less standard error in the theta level. As can be seen in Figure 2 below, Form 3 is the form with the least standard error and the broadest level of theta. The marginal reliability coefficients calculated depending on GPCM for the forms is as follows:

- 0.866 for Form 1,
- 0.885 for Form 2,
- 0.890 for Form 3 and,
- 0.872 for Form 4.

Figure 2

The Test Information Functions of the Forms Based on GPCM



It is possible to see this in Figure 2. Therefore, it can be said that the form which gives the higher information is Form 3 with the statement "*Neither agree nor disagree*." Form 2, which includes the expression "*Undecided*" gives information mostly in the range of -1.0 to 0.5 theta, and Form 3 gives maximum information in the range of -1.0 to 0.0 theta.

Findings as to the item parameters of Form 1-2-3-4 estimated based on GRM

In the fourth sub-problem, the item parameters estimated on the basis of GRM for all forms are given in Table 6.

According to Table 6, the discrimination parameter "a" ranges from 2.09 to 4.85 in Form 1, from 1.85 to 5.38 in Form 2, from 2.00 to 5.95 in Form 3, and from 1.66 to 6.35 in Form 4. The "a" parameter values estimated based on the GRM were found to be higher than GPCM. Also, the range for parameter "a" was highest in Form 4. Like GPCM, in all forms based on GRM, item 3 had the highest and item 8 had the lowest discrimination.

Table 6

The Item Parameters Estimated with GRM

Form 1						Form 2				
	a	*b1*	*b2*	*b3*	*b4*	*a*	*b1*	*b2*	*b3*	*b4*
Item 1	4.10	-1.20	-0.23	0.20	1.30	4.48	-1.23	-0.19	0.30	1.26
Item 2	2.37	-1.89	-0.79	-0.24	1.22	2.43	-1.91	-0.86	-0.23	1.13
Item 3	4.85	-1.34	-0.49	-0.16	0.90	5.38	-1.35	-0.49	-0.01	1.03
Item 4	4.10	-0.98	0.01	0.34	1.26	5.06	-1.01	-0.06	0.40	1.26
Item 5	2.94	-1.49	-0.55	-0.22	1.01	2.47	-1.60	-0.74	-0.26	1.05
Item 6	3.48	-0.91	0.15	0.48	1.17	4.24	-0.85	0.06	0.50	1.12
Item 7	4.00	-1.44	-0.76	-0.48	0.39	4.48	-1.39	-0.75	-0.45	0.47
Item 8	2.09	-1.31	-0.47	-0.09	0.87	1.85	-1.48	-0.49	-0.09	0.96
Item 9	3.05	-1.21	-0.10	0.33	1.20	3.61	-1.10	-0.16	0.36	1.23
Item 10	3.70	-1.30	0.08	0.44	1.45	4.10	-1.28	-0.06	0.45	1.40
Item 11	2.54	-2.02	-0.90	-0.40	1.01	2.80	-1.90	-0.84	-0.26	1.04
Item 12	3.28	-1.62	-0.73	-0.48	0.72	3.15	-1.63	-0.71	-0.39	0.77
Item 13	2.52	-1.81	-0.57	0.00	1.24	2.57	-1.69	-0.64	-0.05	1.32
Item 14	4.37	-1.28	-0.53	-0.24	0.87	4.47	-1.33	-0.46	-0.12	0.85
Mean	3.38	-1.41	-.42	-.03	1.04	3.64	-1.41	-.45	.01	1.06
Median	3.38	-1.320	-.51	-.12	1.09	3.855	-1.37	-.49	-.07	1.08
St. Dev.	.837	.32	.34	.33	.27	1.10	.31	.31	.32	.24
	Form 3					Form 4				
	a	*b1*	*b2*	*b3*	*b4*	*a*	*b1*	*b2*	*b3*	*b4*
Item 1	4.97	-1.15	-0.26	0.31	1.16	4.37	-0.95	-0.24	0.25	0.97
Item 2	2.34	-1.90	-0.90	-0.18	1.28	2.79	-1.35	-0.66	-0.11	0.83

Item 3	5.95	-1.20	-0.43	0.03	1.07	6.35	-1.12	-0.43	0.12	0.87
Item 4	4.67	-0.99	-0.01	0.49	1.30	4.38	-0.79	-0.12	0.40	1.13
Item 5	2.82	-1.49	-0.69	-0.38	0.80	2.50	-1.48	-0.89	-0.35	0.56
Item 6	3.96	-0.81	-0.01	0.51	1.10	3.49	-0.77	-0.05	0.43	0.95
Item 7	4.15	-1.38	-0.73	-0.39	0.48	3.12	-1.31	-0.75	-0.44	0.25
Item 8	2.00	-1.32	-0.42	0.07	1.04	1.66	-1.21	-0.36	0.14	0.84
Item 9	3.96	-1.06	-0.13	0.44	1.27	4.10	-0.88	-0.19	0.36	1.10
Item 10	3.53	-1.22	-0.20	0.46	1.46	3.80	-1.18	-0.20	0.41	1.22
Item 11	2.74	-1.75	-0.91	-0.24	0.98	2.51	-1.51	-0.86	-0.22	0.60
Item 12	3.48	-1.52	-0.75	-0.37	0.79	2.80	-1.35	-0.71	-0.30	0.39
Item 13	2.83	-1.50	-0.58	-0.02	1.26	2.93	-1.34	-0.58	0.02	0.90
Item 14	4.54	-1.30	-0.48	-0.11	0.84	3.23	-1.05	-0.56	-0.17	0.65
Mean	3.71	-1.32	-.46	.04	1.05	3.43	-1.16	-.47	.03	.80
Median	3.74	-1.31	-.45	.00	1.08	3.17	-1.19	-.49	.07	.85
St. Dev.	1.10	.29	.30	.34	.26	1.14	.24	.28	.30	.28

According to Table 6, the standard deviation of parameter “a”, like GPCM, was estimated the highest in Form 4 and the lowest in Form 1. Considering the central tendency measures of parameter “a” in forms, Form 4 had the lowest values, while Form 3 for the mean, and Form 2 for the median had the highest values.

Table 6 shows that the form in which the threshold parameter “b1” deviates most is Form 1, while the form least deviated is Form 4. The tendency measures for this parameter are the highest in Form 4, and the lowest in Form 1 and Form 2. The “b1” threshold parameter ranges from -2.02 to -0.91 in Form 1, from -1.91 to -0.85 in Form 2, from -1.9 to -0.81 in Form 3, and from -1.51 to -0.77 in Form 4.

Table 6 shows that the form with the highest standard deviation is Form 1, and that with the lowest is Form 4 for the parameter “b2”. Parameter “b2” ranges from -0.90 to 0.15 in Form 1, from -0.86 to 0.06 in Form 2, from -0.91 to -0.01 in Form 3, and from -0.89 to -0.05 in Form 4. The smallest central tendency value was calculated in Form 1 as per the median and in Form 4 as per the mean. The form in which this parameter has the highest tendency value is Form 1 as per the mean value and Form 3 as per the median value.

As can be seen in Table 6, the standard deviation values of the parameter “b3” are very close to each other, with the highest standard deviation value in Form 3 and the lowest one in Form 4. Parameter “b3” ranges from -0.48 to 0.48 in Form 1, from -0.45 to 0.5 in Form 2, from -0.39 to 0.51 in Form 3, and from -0.44 to 0.43 in Form 4. The highest mean of this parameter was obtained from Form 3, and the highest median was obtained from Form 4. On the other hand, the lowest central tendency values were obtained from Form 1.

According to Table 6, the standard deviation values of the threshold parameter “b4” are also close to each other, with the highest deviation in Form 4 and the least deviation in Form 2. The parameter values range from 0.39 to 1.45 in Form 1, from 0.47 to 1.4 in Form 2, from 0.48 to 1.46 in Form 3, and from 0.25 to 1.22 in Form 4. The highest mean of this parameter was obtained from Form 2, and the highest median was obtained from Form 1, while the lowest central tendency values were obtained from Form 4. Commonly in both models, in Form 4, parameter “b1” had the highest value, and parameter “b4” had the lowest value.

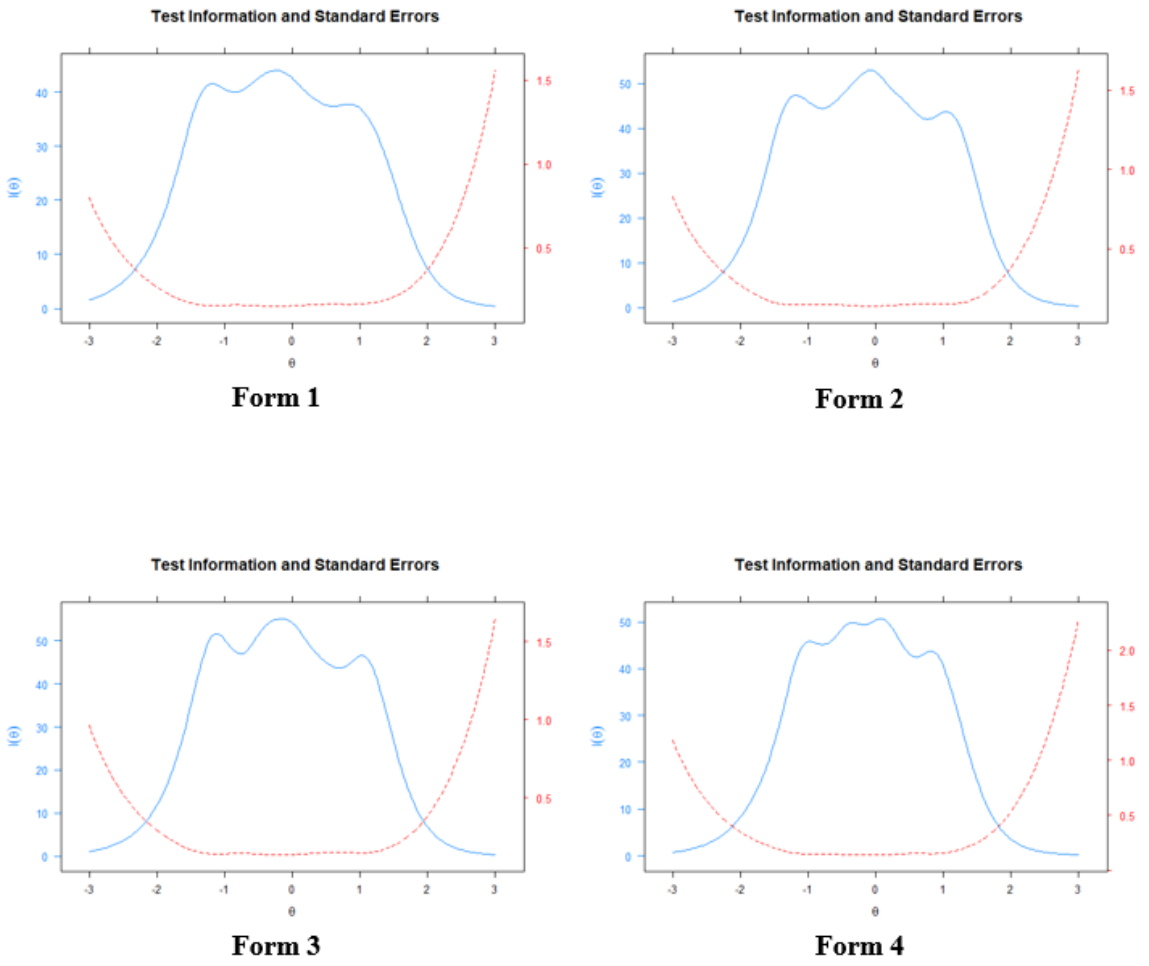
Findings as to the reliability coefficients of Form 1-2-3-4 estimated based on GRM

In the fifth sub-problem, the estimated reliabilities based on GRM were examined. Figure 3 shows that the test information functions provide a high level of information in a wider theta range in the GRM. The marginal reliability calculated based on GRM for the forms is as follows:

- 0.934 for Form 1,
- 0.937 for Form 2,
- 0.939 for Form 3 and,
- 0.935 for Form 4.

Figure 3

The Test Information Functions of the Forms Based on GRM



The marginal reliability obtained from the forms provides information about the information of the tests. According to the results obtained, although the reliability estimated from the forms provides similar results, the form that gives information at maximum level is Form 3, and the form that gives information at minimum level is Form 1. The test information functions in Figure 3 support these findings on reliability. As per this model, the forms are tended to give information at maximum level for wider theta range. It can be said that, compared to GPCM, the parameters estimated based on the GRM give more information.

Findings on the relationship between the item parameters estimated based on GPCM and GRM for the Form 1-2-3-4

The Spearman correlation coefficients were calculated, and were given in Table 7 for the relationships between item parameters estimated from different forms under the same model, and between item parameters estimated based on two models of the same form to examine the compatibility of the item parameters depending on GPCM and GRM with each other in the sixth sub-problem.

Table 7

The Relationship Between the Item Parameters Estimated on the Basis of GRM and GPCM

	a	b1	b2	b3	b4
The relationship between the item parameters based on GPCM					
Form 4 – Form 3	.84*	.79*	.81*	.95*	.82*
Form 4 – Form 2	.82*	.69*	.84*	.85*	.78*
Form 4 – Form 1	.68*	.89*	.84*	.84*	.80*
Form 2 – Form 3	.93*	.95*	.94*	.95*	.85*
Form 1 – Form 3	.91*	.96*	.94*	.91*	.90*
Form 1 – Form 2	.92*	.89*	.98*	.95*	.94*
The relationship between the item parameters based on GRM					
Form 4 – Form 3	.90*	.94*	.94*	.99*	.89*
Form 4 – Form 2	.88*	.93*	.92*	.98*	.82*
Form 4 – Form 1	.80*	.94*	.91*	.93*	.82*
Form 2 – Form 3	.96*	.98*	.96*	.98*	.87*
Form 1 – Form 3	.94*	.96*	.98*	.93*	.90*
Form 1 – Form 2	.95*	.97*	.98*	.95*	.98*
The relationship between the item parameters based on both models					
Form 1 $_{\text{GPCM-GRM}}$	.98*	.95*	.87*	.85*	.97*
Form 2 $_{\text{GPCM-GRM}}$	.96*	.99*	.86*	.88*	.99*
Form 3 $_{\text{GPCM-GRM}}$	.97*	.97*	.87*	.93*	.97*
Form 4 $_{\text{GPCM-GRM}}$	.96*	.76*	.85*	.98*	.89*

*$p < .01$

Table 7 shows that when the correlations between the estimated item parameters of the forms are examined depending on GPCM, all values are statistically significant, and almost all are at a high level. Only two correlation coefficients were calculated less than 0.70. One of them is the coefficient for parameter "a", which was estimated from Form 1 and Form 4, and the other is the coefficient for parameter b1, which was estimated from Form 2 and Form 4. These two values are 0.68 and 0.69, respectively, yet both indicate a significant correlation at the upper intermediate level.

As can be seen in Table 7, all forms are 0.80 and more among the item parameters estimated based on GRM, and have significant correlations. In both models, the highest correlation coefficient for the discrimination parameter "a" is between Form 2 and Form 3, while the lowest coefficient is between Form 4 and Form 1. Therefore, in case the middle categories were "*Undecided*" (Form 2) and "*Neither agree nor disagree*" (Form 3), the discrimination indices of the items yielded similar results. The discrimination indices for the category "*I have no idea*" (Form 1) and the scale form at grading-level (Form 4) differed. The relationship between the threshold parameters is acceptable in both models.

In Table 7, the relationship between Form 3 and Form 4 shows that the correlations of the parameters in the forms are higher in the GRM model. The threshold parameter with the highest similarity between the parameters of these forms estimated based on GPCM is "b3" with 0.95, and the lowest threshold parameter is "b1" with 0.79. The threshold parameters represent the trait level necessary to respond above the j threshold with 0.50 probability (Embretson & Reise, 2000). Because the sample to which the forms were administered was the same in this study, the differentiation in the theta level seen in the forms pointed to the differentiation of individuals in perceiving the forms. Therefore, this shows that the "*Neither agree nor disagree*" and category "*3*" for Form 3 and Form 4, the similarity between the skill levels required to answer in the upper category with a probability of 0.5 from the middle category is 95%. Similarly, in the "b1" threshold parameter, the similarity between the skill levels required to answer in the upper category with a probability of 0.5 from the lowest category marked is. 79%. The most similar parameter did not change (r=0.99 in category b3) in the GRM model, but the parameter with the lowest relationship was "b4" with 0.89. Accordingly, the similarity between the skill levels required to answer in the highest categories based on the GRM had relatively lower correlation than other categories.

Table 7 shows that the parameter with the lowest correlation between the parameters of Form 4 and Form 2 was in "b1" parameter (0.69), and the highest relationship was in "b3" parameter (0.85) according to GPCM. In GRM, similar to GPCM, these correlations were found to be the highest in "b3" parameter (0.98), and the lowest in "b4" parameter (0.82). Regarding the correlations between Form 4 and Form 1, the lowest correlations were between the parameters obtained from these forms. The lowest correlation between the estimated parameters for these forms was in parameter "a" (0.68 for GPCM and 0.80 for GRM), and the highest correlation was in parameter "b1" (0.89 for GPCM and 0.94 for GRM). The parameters based on the two models for Form 2 and Form 3, Form 1 and Form 2, and Form 1 and Form 3, with varying mid-point name only, showed differences. According to these correlation coefficients, the fact that only the name of the middle value differed in the forms caused the other categories to be understood differently for these forms.

Findings on the differences between the item parameters estimated based on GPCM for the Form 1-2-3-4

In the solution of the seventh sub-problem, whether the item parameters estimated based on GPCM differed by the forms was tested first using the Friedman test and then the Wilcoxon test, and the test results were given in Table 8.

Table 8

The Investigation of the Difference of the Parameters Estimated from GPCM Between the Forms

Parameter	Form	Mean Rank	Chi-squared	df	p	Significant Difference
a	Form 1	1.93	11.743	3	.008	Form 1 - Form 2
	Form 2	2.86				Form 1 - Form 3
	Form 3	3.29				Form 2 - Form 4
	Form 4	1.93				Form 3 - Form 4
b1	Form 1	1.50	31.542	3	.000	Form 1 - Form 3
	Form 2	1.64				Form 2 - Form 3
	Form 3	2.86				Form 1 - Form 4
	Form 4	4.00				Form 2 - Form 4
						Form 3 - Form 4
b2	Form 1	3.93	26.486	3	.000	Form 1 - Form 2
	Form 2	2.50				Form 1 - Form 3
	Form 3	2.00				Form 1 - Form 4
	Form 4	1.57				
b3	Form 1	1.29	20.828	3	.000	Form 1 - Form 2
	Form 2	2.36				Form 1 - Form 3
	Form 3	3.36				Form 2 - Form 3

	Form 4	3.00				Form 1 - Form 4
b4	Form 1	2.86	25.543	3	.000	Form 1 - Form 4
	Form 2	3.00				Form 2 - Form 4
	Form 3	3.14				Form 3 - Form 4
	Form 4	1.00				

Table 8 shows whether the median values of the parameters estimated depending on GPCM differed by forms. According to the Friedman test results, both the discrimination and threshold parameters showed a significant difference from one form to another ($p < .01$). The differences between paired combinations of the forms were examined with Wilcoxon test. Paired comparisons results showed that the discrimination parameters "a" estimated based on the GPCM differed significantly between Form 4 and Form 3, Form 4 and Form 2, Form 1 and Form 2, Form 1 and Form 3. When the middle value was "*I have no idea*" (Form 1) and "*3*" (Form 4), the discrimination parameters in the forms had similar values, and these values were relatively lower than other values. Furthermore, discrimination parameters obtained for middle categories of "*Neither agree nor disagree*" (Form 3) and "*Undecided*" (Form 2) had similar values in the forms.

As seen in Table 8, the threshold parameters estimated based on the GPCM also differed by the forms ($p < .01$). These differences were tested on paired combinations of the forms. The values of all threshold parameters obtained from Form 1 (*b1, b2, b3, b4*) and all threshold parameters obtained from Form 4 differed significantly. This finding can be an indicator that presenting the middle value as "*I have no idea*" or presenting the categories at grading-level numerically may create a difference in the perceptions of individuals.

Although Form 4 differed from other forms in terms of containing numerical values, it did not differ from Form 2 and Form 3 in terms of "b2" and "b3" parameters. However, it differed significantly from these forms in terms of the terminal/extreme threshold parameters "b1" and "b4". In this case, it can be interpreted that Form 2 with the middle category "*Undecided*" and the Form 4 with the middle value "*3*" differ from each other in terms of the marking of the extreme categories. Similarly, the Form 3 with the middle category "*Neither agree nor disagree*" and Form 4 differed significantly from each other in the way of marking the highest and lowest categories.

Although Form 4 has only numerical categories, and seems to differ from other forms, there were significant differences in terms of the parameters among other forms whose only middle values differed, also. In particular, parameters "b1", "b2", "b3" estimated from Form 1 with the category "*I have no idea*" and Form 3 with the category "*Neither agree nor disagree*" differed significantly. This differentiation can be expressed as the change in the name of the middle category in the forms affects the marking of the extreme category "*Strongly Disagree*".

Table 8 shows that "b2" and "b3", among the threshold parameters of Form 1 with the middle category "*I have no idea*" and Form 2 with the middle category "*Undecided*", differed significantly from one

form to another. This differentiation means that the marking level of the middle categories of these forms changed only with the change of the name of the middle category, and individuals were more likely to prefer lower categories for Form 1.

For Form 2 with a middle category "*Undecided*" and Form 3 with a middle category "*Neither agree nor disagree*", the threshold parameters of "b1" and "b3" differed significantly between forms. This differentiation indicated that the marking levels of the middle categories of these forms altered with the change in the middle category name. Both threshold parameters had higher values in Form 3. It can be interpreted that the change in the middle category name increased the possibility of preferring the lowest category, "*Strongly disagree*", and the middle category, "*Neither agree nor disagree*".

Findings on the differences between the item parameters estimated based on GRM for the Form 1-2-3-4

In the solution of the eighth sub-problem, whether the item parameters estimated based on GRM differed by the forms was tested first using the Friedman test and then Wilcoxon test, and the test results were given in Table 9.

Table 9

The Investigation of the Difference of the Parameters Estimated from GRM Between the Forms

Parameter	Form	Mean Rank	Chi-squared	df	p	Significant Difference
a	Form 1	1.93	5.743	3	.125	
	Form 2	2.86				-
	Form 3	2.93				
	Form 4	2.29				
b1	Form 1	1.89	31.457	3	.000	Form 1 – Form 3
	Form 2	1.43				Form 1 – Form 4
	Form 3	2.68				Form 2 – Form 3
	Form 4	4.00				Form 2 – Form 4
						Form 3 – Form 4
b2	Form 1	2.89	1.971	3	.533	-
	Form 2	2.54				
	Form 3	2.32				

	Form 4	2.25				
b3	Form 1	1.54	14.314	3	.002	Form 1 - Form 3
	Form 2	2.32				Form 1 – Form 2
	Form 3	3.29				Form 1 – Form 4
	Form 4	2.86				Form 2 – Form 3
b4	Form 1	2.75	26.309	3	.000	Form 1 – Form 4
	Form 2	3.04				Form 2 – Form 4
	Form 3	3.21				Form 3 – Form 4
	Form 4	1.00				

According to Table 9, discrimination parameter "a" estimated based on the GRM does not show a significant difference between forms ($p > .01$). Thus, it can be said that the parameters "a" estimated from all forms are similar. The threshold parameters, except for parameter "b2", showed significant differences between the forms ($p < .01$). In paired comparisons, except for "b2", all threshold parameters of Form 1 and Form 4 significantly differed from each other. While the lower extreme category in Form 4 with numerical categories was preferred more than Form 1, the upper extreme categories were preferred more in Form 1. However, for Form 2 with a middle category "*Undecided*" and Form 3 with the middle category name of "*Neither agree nor disagree*", the threshold parameters of "b1" and "b3" differed significantly between forms. This differentiation is similar to the findings from the GPCM.

As seen in Table 9, similar to the findings obtained from the GPCM, the extreme threshold parameters "b1" and "b4" parameters differed significantly between Form 4 and other forms. Furthermore, there was a significant difference between the "b1" extreme threshold parameters estimated from Form 3 and obtained from Form 2. Unlike GPCM, only "b3" threshold parameters differed significantly from Form 1 with the middle category "*I have no idea*" to Form 2 with the middle category "*Undecided*". Only with the change of the middle category name did the levels of marking the middle categories of these forms change, and there was a tendency in Form 2 to mark more the middle levels. For parameters estimated from Form 1 with the middle category "*I have no idea*" and Form 3 with the middle category "*Neither agree nor disagree*", unlike GPCM, only parameters "b1" and "b2" differed significantly. Therefore, the forms differed in general according to both GPCM and GRM, and the differentiation of the middle category names caused significant differences in the marking levels of the other categories.

RESULTS, DISCUSSIONS, AND SUGGESTIONS

Researchers may use a scale with a single or double category. In Likert-type scales without mid-point, because participants are forced to select a degree of disagreement or agreement (Chyung, et al., 2017; Johns, 2005), these scales are used as forced-choice scales (Chyung, et al., 2017). However, the use of a middle category can contribute to the improvement of reliability, and make a difference for individuals in the measurement of an affective trait (Adelson & McCoach, 2010; Croasmun & Ostrom, 2011; O'Muircheartaigh, Krosnick, & Helic, 2000). On a scale with no middle option, respondents must make a decision to express their opinion. In this case, it will guide the participants who do not have an opinion on a subject, or are unsure of their opinions, to choose only one option. The answers received from individuals who are directed to make a choice will not reflect reality, and people who are undecided or have no idea about that question may not want to fill in the scale. Even those who do not have an idea about how to answer the items may not be right to respond to those items. Having the middle option will partially prevent these referrals, and allow the items to be more distinctive. In this study, 5-Likert-type scoring situation was discussed. Forced choice scales can also be examined in further research. Just as offering the options to individuals with different priorities may create a difference in perceptions, offering the categories in the scales to individuals in different formats may cause differences in perceptions as well. Varied labeling in the middle categories does not give clear information about whether the participants are actually unfamiliar with the subject, or whether they are actually undecided. Furthermore, participants may use the middle category as an escape in marking (Chyung, et al. 2017; Kulas & Stachowski, 2009). To examine the differentiation in such perceptions, the effects of differences in category labels in Likert-type scales on individuals' perceptions were examined considering polytomous IRT models. In many studies examining model fit, the GRM was found to be more compatible in the comparison of GRM and GPCM (Büyükkıdık & Atar, 2018; Schneider, et al., 2020; Sözer & Kahraman, 2021; Yaşar & Aybek, 2019;). The reason for this can be expressed as the fact that GRM is based on the gradual presentation of items in the estimation (Embretson & Reise, 2000; Hays, Morales, & Reise, 2000). However, it is difficult to distinguish between these two models as their item parameters are similar (Dai et al., 2021). For this purpose, the Mathematical Attitude Scale (MAS) was used with various category names. This study used GPCM and GRM as base among unidimensional polytomous IRT models. Considering both models for four different forms, the forms showed better fit in different models. The result of this study showed that Form 1 with the middle category "*I have no idea*" and Form 2 with the expression "*Undecided*" in the middle category showed better fit to GRM. However, Form 3 with the middle category name of "*Neither agree nor disagree*" and Form 4 with numerical property showed better fit to GPCM. Therefore, examinations were made considering all estimations for both models throughout the research.

Regarding the item parameters estimated in accordance with GPCM and GRM on all forms, the same items for both models had the highest discrimination "a" parameter (item 3) and the lowest a parameter (item 8). When the item parameters in the forms were estimated based on two different IRT models, Form 4, unlike other forms, had a wide range of parameters "a", and had a narrow range of "b1" values. These results can be an indication that the categories in Form 4 differentiate the students' perceptions.

Marginal reliability provides an indication of the overall consistency of test scores in measuring the observed score or underlying trait generated from the IRT model (Andersson & Xin, 2018). Therefore, if the feature to be measured has different categories, examining the marginal reliability shows that

the scores obtained for the structure are reliable. According to marginal reliability coefficients and test information functions, the most informative and reliable form depending on GPCM was Form 3 with the expression "*Neither agree nor disagree*" in the middle category. Compared to GPCM, the reliability of the forms being higher in GRM, and the use of the models suitable for the forms only would be restrictive for the reliability and the estimations. Regarding the model alignments of GRM and GPCM, although there are studies suggesting that the fit of the two models shows similarity (Schneider, et al., 2020), there are many studies suggesting that GRM has better fit than GPCM (Büyükkıdık & Atar, 2018; Sischka, et al., 2020; Sözer & Kahraman, 2021; Yaşar & Aybek, 2019;). In this sense, it can be said that the results of the present study are in line with the literature.

It is another important result of the study that all correlation coefficients between the estimated item parameters based on both models are statistically significant. Nevertheless, it is remarkable that the correlations between the parameters obtained from the forms based on GRM are more similar to each other than those obtained by GPCM. The highest correlation coefficient for the discrimination parameters is in Form 2 with the middle category name "*Undecided*" and Form 3 with the statement of "*Neither agree nor disagree*", while the lowest coefficient is in the Form 4 which is the numerical scale and Form 1 with the middle category "*I have no idea*". The correlations between the item parameters estimated depending on GRM and GPCM of the forms showed that the calculated correlation values were high and significant. Therefore, the parameters estimated based on different IRT models yielded very close results.

In the present study, the parameters estimated from the forms depending on the same model were compared so as to examine whether alteration of the middle category name changed the perceptions of individuals and therefore the test-item parameters at a significant level. While the discrimination parameter differed by the forms in GPCM, there was no significant difference between the forms in GRM. Therefore, while the discrimination parameters estimated based on GRM on the forms with different middle category name gave similar results, they showed significant differences from one form to another in GPCM. So, it can be interpreted that GRM is more able to tolerate label/name differences with middle category, or is not affected much by middle category names.

In terms of threshold parameters, "b1", "b3", and "b4", estimated from both IRT models differed by the forms, and this difference was found between similar forms. However, more parameters estimated based on GPCM showed differences. According to both IRT models, the lowest extreme category in Form 4 (numerical scale) was preferred more than other forms. It was concluded that this difference was at a significant level with the forms those are with the middle category name "*I have no idea*" and "*Undecided*". However, the difference in the form with middle category "*Neither agree nor disagree*" was less and insignificant. Similar to the form with numerical categories, the middle value of marking the extreme category differed significantly from marking the forms "*I have no idea*" and "*Undecided*". Based on this finding, providing the categories numerically or labeling the middle category as "*Neither agree nor disagree*" directs the participants' perceptions to mark the lower category. The least preferred form of the highest extreme category is Form 4, which has numerical categories, and this differs significantly from the level of preference in other forms. The form in which the highest extreme category is preferred in other forms is the middle category of "*Neither agree nor disagree*". However, the level of preference for this form, and forms "*I have no idea*" and "*Undecided*" is not significant. Consequently, the participants avoid giving the highest numerical value, and show similar attitudes in other forms. As regard to the preference of the middle categories, the forms "*Undecided*" and "*I have no idea*" had higher preference.

All these findings obtained from the study suggest that the middle category is an important factor in Likert-type scales. Previous studies indicated that if the middle category is single or double, the response style varies (Chyung, et al., 2017; Moors, 2008). It can be questionable whether there is really a midpoint of view between disagreement and agreement, or whether it should be treated as a lack of view. Because the use of the midpoint with different labels can evoke these senses, (Chyung, et al., 2017) differences may occur in the answers. For the Likert scale to be an interval scale, the distances between sequential points on the scale should be the same, that is, the distance between consecutive two categories must be equal (Wakita, Ueshima, & Noguchi, 2012). While consecutive categories can be perceived as numerically equivalent to the same value for Form 4, which represents numerically given categories, it can be difficult to control the distances between consecutive categories in verbal expressions of other forms. This can be an indication of Form 4 taking different values from other forms.

Future studies may use IRT based models, and use the rating scale categories with numbers along with category labels instead of using them alone while preparing forms that can reflect the interests, attitudes, tendencies, or emotions of individuals during scale development. During labeling of categories, the labels can be determined by examining the fit between the data set and model with the use of GRM or GPCM among IRT models particularly. In the current study, the students received the forms in the same order for a certain period of time due to the pandemic. Performance of application in the same order and in the order determined by the researchers can be considered the limitation of this study. Future studies may examine the model fit when forms are distributed to the participants in different orders as well as the effect of the middle category on perception.

REFERENCES

Albaum, G. (1997). The Likert scale revisited: An alternate version. *Journal of the Market Research Society, 39*(2), 331-342. https://doi.org/10.1177/147078539703900202

Adelson, J. L., & McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-type scale. *Educational and Psychological Measurement, 70*(5), 796–807. https://doi.org/10.1177/0013164410366694

Andersson. B., & Xin, T. (2018). Large sample confidence intervals for item response theory reliability coefficients. *Educational Psychological Measurement*, *78*(1), 32-45. https://doi.org/10.1177/0013164417713570

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics, 45*, 966-987. https://doi.org/10.1080/00140130210166951

Blumberg, H. H., DeSoto, C. B. & Kuethe, J. L. (1966). Evaluation of rating scale formats. *Personnel Psychology, 19*, 243-259. https://doi.org/10.1111/j.1744-6570.1966.tb00301.x

Büyükkıdık, S., & Atar, H. (2018). Çok kategorili item tepki kuramı modellerinin örneklem büyüklüğü açısından incelenmesi [Examining multi-category item response theory models in terms of sample size]. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *38*(2), 663-692. https://doi.org/10.17152/gefad.334608

Bartolucci, F., Bacci, S., & Gnaldi, M. (2015). *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Boca Raton, FL: Chapman and Hall/CRC.

Carle, A. C., Jaffee, D., Vaughan, N. W., & Eder, D. (2009). Psychometric properties of three new national survey of student engagement based engagement scales: An item response theory analysis. *Research in Higher Education*, *50*, 775-794. https://doi.org/10.1007/s11162-009-9141-z

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the Likert scale. *Performance Improvement*, *56*(10), 15-23. https://doi.org/10.1002/pfi.21727

Cordier, R., Munro, N., Wilkes-Gillan, S., Speyer, R., Parsons, L., & Joosten, A. (2019). Applying Item Response Theory (IRT) modeling to an observational measure of childhood pragmatics: The pragmatics observational measure-2. *Frontiers in Psychology*, *10*, 408. https://doi.org/10.3389/fpsyg.2019.00408

Croasmun, J. T., & Ostrom, L. (2011). Using Likert-type scales in the social sciences. *Journal of adult education*, *40*(1), 19-22. Retrieved from https://eric.ed.gov/?id=EJ961998

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021, September). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. In *Frontiers in Education* (Vol. 6, p. 721963). Frontiers Media SA. https://doi.org/10.3389/feduc.2021.721963

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all category defined and end-defined Likert formats. *Educational & Psychological Measurement, 44*, 61-66. https://doi.org/10.1177/0013164484441006

Dunkel, A. (2015). Visualizing the perceived environment using crowdsourced photo geodata. *Landscape and urban planning*, *142*, 173-186. https://doi.org/10.1016/j.landurbplan.2015.02.022

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Psychology Press.

Erkuş, A. (2012). *A measurement and scale development in Psychology I: Basic concepts and processes.* Ankara: Pegem Academy Publishing (In Turkish).

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education, 32*(2), 77–96. https://doi.org/10.1080/08957347.2019.1577243

Finn, R. H. (1972). Effects of some variations of rating scale characteristics on the means and reliabilities of ratings. *Educational & Psychological Measurement, 32*, 255-265. https://doi.org/10.1177/001316447203200203

Gibson, J.L., Ivancevich, J.M., James H., & Donnely Jr. (1996), *Organizational behavior structure*, Process. 9th Edition, Chicago: Irwin.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer Nijhoff Publishing.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical care, 38*(9), 28-42. https://doi.org/10.1097%2F00005650-200009002-00007

Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, *7*(1706), 1-15. https://doi.org/10.3389/fpsyg.2016.01706

Hulın, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Jacko, E. J., & Huck, S. W. (1974, April). The Effect of varying the response format on the statistical characteristics of the Alpert-Haber Achievement Anxiety Test. Paper presented at the *Annual Meeting of the American Educational Research Association* (59th, Chicago, Illinois).

Jin, K. Y., & Wang, W. C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement, 51*, 178–200. https://doi.org/10.1111/jedm.12041

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters. *Applied Psychological Measurement, 36*(5), 399–419. https://doi.org/10.1177/0146621612446170

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal, 6*(2),151-162. https://doi.org/10.32614/RJ-2014-031

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., ... & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International journal of nursing studies*, *48*(6), 661-671. https://doi.org/10.1016/j.ijnurstu.2011.01.016

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science, 37*, 941–964. https://doi.org/10.2307/2111580

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality, 43*(3), 489-493. https://doi.org/10.1016/j.jrp.2008.12.005

Lange, T., Schmitt, J., Kopkow, C., Rataj, E., Günther, K. P., & Lützner, J. (2017). What do patients expect from total knee arthroplasty? A Delphi consensus study on patient treatment goals. *The Journal of arthroplasty*, *32*(7), 2093-2099. https://doi.org/10.1016/j.arth.2017.01.053

Mendiburu, F. D. (2021). Agricolae: Statistical Procedures for Agricultural Research. 2017. *R package version*, 1-1.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & quantity*, *42*, 779-794. https://doi.org/10.1007/s11135-006-9067-x

Muraki, E. (1992). *A generalized partial credit model: Application of an em algorithm.* ETS research report-1, i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik Likert tipi ölçek ile metrik ölçeğin Item ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi*. *[Examining the Item and scale properties of Likert-type scale and metric scale for measuring the same attitude according*

to classical test theory and latent trait theory.] Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Yayımlanmamış Doktora Tezi, Ankara.

Newstead, S. E. & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational & Psychological Measurement, 49*, 33-43. https://doi.org/10.1177/0013164489491004

OECD (2021). *PISA 2018 Technical Report*. Paris: Organization for Economic Cooperation and Development (OECD). https://www.oecd.org/pisa/data/pisa2018technicalreport/

Ostini, R. & Nering, M. L. (2006). *Polytomous item response theory models*. California: Sage.

Ogle, D. H., Wheeler, P. & Dinno, A. (2021). *FSA: Fisheries Stock Analysis*. R package version 0.9.0, Retrieved from https://github.com/droglenc/FSA.

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (2000). *Middle alternatives, acquiescence, and the quality of questionnaire data*. The Center for Advanced Study in the Behavioral Sciences. Retrieved from:https://www.academia.edu/18408388/Middle_Alternatives_Acquiescence_and_the_Quality_Questionnaire_Data?bulkDownload=thisPaper-topRelated-sameAuthor-citingThis-citedByThis-secondOrderCitations&from=cover_page

Pomerantz, J. R. (2003). *Perception: Overview*. In: Lynn Nadel (Ed.), Encyclopedia of Cognitive Science, Vol. 3, London: Nature Publishing Group, pp. 527–537.

R Development Core Team. (2013). *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.Rproject.org

Rajamanickam M (2007) *Modern general psychology thoroughly revised and expanded*, 2nd edn. Concept Publishing Company, New Delhi, p.330

Robitzsch, A., Kiefer, T., & Wu, M. (2021). *TAM: Test Analysis Modules*. R package version 3.7 16, Retrieved from: https://CRAN.R-project.org/package=TAM

Qiong, O. U. (2017). A brief introduction to perception. *Studies in Literature and Language*, *15*(4), 18 - 28. https://doi.org/10.3968/10055

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No.17.

Schneider, L., Chalmers, R. P., Debelak, R. &. Merkle, E. C. (2020) Model selection of nested and non nested item response models using vuong tests, *Multivariate Behavioral Research, 55*(5), 664-684, https://doi.org/10.1080/00273171.2019.1664280

Sischka, P. E., Costa, A. P., Steffgen, G., & Schmidt, A. F. (2020). The WHO-5 well-being index validation based on item response theory and the analysis of measurement invariance across 35 countries. *Journal of Affective Disorders Reports*, *1*, 100020. https://doi.org/10.1016/j.jadr.2020.100020

Sözer, E., & Kahraman, N. (2021). Investigation of psychometric properties of Likert items with the same response categories using polytomous item response theory models. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 129-146. https://doi.org/10.21031/epod.819927

Steiner, M., & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), 2521. https://doi.org/10.21105/joss.02521

Sung, H. J., & Kang, T. (2006, April). Choosing a polytomous IRT model using Bayesian model selection methods. In *National Council on Measurement in Education Annual Meeting* (pp. 1-36).

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics.* USA: Pearson Education Inc.

Yaşar, M., & Aybek, E. C. (2019). Üniversite öğrencileri için bir yılmazlık ölçeğinin geliştirilmesi: Item tepki kuramı temelinde geçerlilik ve güvenilirlik çalışması [Development of a resilience scale for university students: A validity and reliability study based on item response theory]. *İlköğretim Online, 18*(4), 1687 -1699. Retrieved from https://ilkogretim-online.org/fulltext/218-1597121020.pdf?1618815938

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement, 72*, 533-546. https://doi.org/10.1177/0013164411431162

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76*(2), 304-324. https://doi.org/10.1177/0013164415591848

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four-point Likert-type response scales. *Educational & Psychological Measurement, 47*, 27-35. https://doi.org/10.1177/0013164487471003

Author Contributions

First author planned and modeled the study. Second author co-wrote the paper with third author and first author who were involved in the collection of the data. Second author performed the data analysis of the study and contributed to the interpretation of the results. Third author contributed to the literature review and discussion section.

Conflict of Interest

No potential conflict of interest was declared by the author.

Supporting Individuals or Organizations

No grants were received from any public, private or non-profit organizations for this research.

Ethical Approval and Participant Consent

Ethics committee permission for this study was obtained from İnönü University Scientific Research and Ethics Committee with the decision dated 02.07.2021 and numbered 2021/13-19.

Plagiarism Statement

Similarity rates of this article was scanned by the iThenticate software. No plagiarism detected.

Availability of Data and Materials

Not applicable.

Acknowledgements

No acknowledgements.