International Journal of Assessment Tools in Education
2024, Vol. 11, No. 3, 524–536
https://doi.org/10.21449/ijate.1400157
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

How many grades of response categories does the commitment to the profession of medicine scale provide the most information?

Murat Tekin[1], **Çetin Toraman**[1*], **Ayşen Melek Aytuğ Koşan**[1]

[1]Çanakkale Onsekiz Mart University, Faculty of Medicine, Department of Medical Education, Çanakkale, Türkiye

ARTICLE HISTORY

Received: Dec. 04, 2023

Accepted: July 24, 2024

Keywords:
Likert scale,
Response set,
Item response theory,
Medical student.

Abstract: In the present study, we examined the psychometric properties of the data obtained from the Commitment to Profession of Medicine Scale (CPMS) with 4-point, 5-point, 6-point, and 7-point response sets based on Item Response Theory (IRT). A total of 2150 medical students from 16 different universities participated in the study. The participants were divided into four groups consisting of 560, 544, 502, and 544 medical students. The first group (n=560) was assigned four-point, the second group (n=544) five-point, the third group (n=502) six-point, and the fourth group (n=544) seven-point Likert forms. We used R statistical software to analyze the data. The results of item calibrations conducted with the Graded Response Model (GRM) were analyzed. The results show that the eigenvalue increased from 4-point to 7-point. Similarly, the explained variance percentage and the scale's reliability increased gradually from 4-point to 7-point. The explained variance, reliability level, and eigenvalue were very close in the 5-point and 6-point forms.

1. INTRODUCTION

Scales are used to collect data in many scientific fields. Scales can be configured with the Thurstone scaling technique (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017; Torgerson, 1958), Guttman scaling technique (Anastasi & Urbina, 1997; Dunn- Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017), and the Likert scaling technique (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Price, 2017). In Likert-type scales, mainly used to measure thoughts, beliefs, and attitudes, the participants' level of agreement in the statements given is measured by grading with the Likert scaling technique (Anastasi & Urbina, 1997; DeVellis, 2003). Likert scales are very popular in use as they are easy to configure. Likert scales are widely used in social sciences and educational research (Joshi et al., 2015).

When taking the participants' answers, distances between each choice (answer option) are assumed to be equal in Likert scales. This is because Likert (1932) suggests that the "distance between response categories is assumed to be equal". Response set may broadly include five points to a statement: (1) Strongly disagree, (2) Disagree, (3) Neither agree nor disagree, (4) Agree, and (5) Strongly agree (Anastasi & Urbina, 1997). Additionally, they may include six

*CONTACT: Çetin TORAMAN ✉ toramanacademic@gmail.com Çanakkale Onsekiz Mart University, Faculty of Medicine, Department of Medical Education, Çanakkale, Türkiye

e-ISSN: 2148-7456

points to a statement as well: (1) Disagree very strongly, (2) Disagree strongly, (3) Disagree, (4) Agree, (5) Agree strongly, and (6) Agree very strongly (DeVellis, 2003). Likert types can have response sets with or without the 'neutral' option. Although there are grade suggestions such as "neither agree nor disagree" or "equally agree and disagree" for the neutral point, discussions regarding this neutral expression continue (DeVellis, 2003). Although the average scores on the Likert scale were not substantially affected by the inclusion or exclusion of a "Neutral" option, significant variations emerge when combining neighboring categories, such as the proportion of respondents who "Agree or Strongly Agree." This suggests that the presence or absence of a neutral category can lead to considerably different interpretations. Respondents may find the neutral option valuable, and its removal could result in misleading conclusions, especially when analyzing individual items. In contexts where the scale is used for quality enhancement or progress tracking, including a neutral option may provide a more accurate reflection of shifts in perception. (Mariano et al., 2024). It is important to determine whether Likert scales obtain data at the ordinal or interval scale through response categories. While some researchers (Jamieson, 2004; Stevens, 1946; Thomas, 1982) claim that the data obtained from Likert scales are at the ordinal scale level, some others (Norman, 2010) assert that it can be accepted at an interval scale level and parametric analyzes can be used in line with this assumption. Some studies suggest that by increasing the number of grades in the answer set, the obtained data will be normally distributed and set to an interval scale level (Wu & Leung, 2017). Several studies have been conducted on the descriptive statistics of data obtained from Likert scales with varying response categories using 5, 7, and 10-degree response sets. It was determined that the scale mean with 10 response categories tended to be lower than the scale mean with 5 or 7 response categories. The scales offered very similar values in terms of skewness and kurtosis (Dawes, 2008). In another study conducted with Likert scales with different response categories (4, 5, 6, and 11 response categories), no major differences could be determined between the mean, standard deviation, item correlations, Cronbach Alpha value, and factor loadings of the data obtained. The skewness and kurtosis of the data obtained from the scale with the most response categories (11 degrees) decreased and approached normal distribution (Leung, 2011). In another study, the data obtained from the Likert scale, prepared in different forms as 5, 7, 9, and 11 response categories, were compared in terms of mean, standard deviation, skewness, and kurtosis. The increase in response categories caused the mean to decrease. According to the skewness value, the closest scale to the normal distribution is the 5-degree scale. According to the kurtosis value, the scale closest to the normal distribution is the 11-category scale (Bora, 2013).

Likert-type scales have been the subject of extensive research studies on:

- The effect of the number of categories in the response set on the alpha coefficient (Aiken, 1983; Chang, 1994; Leung, 2011; Wong et al., 1993),
- Its effect on test-retest reliability level (Preston & Colman, 2000),
- How many grades an answer set should have (Champney & Marshall, 1939),
- How the number of grades in the response set affects the arithmetic means and distribution measures (standard deviation, kurtosis, skewness) of the data obtained (Bora, 2013; Dawes, 2008; Leung, 2011),
- Its effects on the normal distribution (Leung, 2011),
- Participants' perceptions of variables in the answer set (Adelson & McCoach, 2010),
- How the number of grades in a response set affects item parameters based on item response theory (IRT) (Aybek & Toraman, 2022; Wakita et al., 2012).

In summary, agreement categories of relevant Likert model scales were examined based on reliability, covariance matrices, descriptive statistics, the ability to distinguish the neutral option in the response set, and the effect on factor loads in terms of classical test theory (CTT). Aybek and Toraman, (2022) and Wakita et al., (2012) examine the effect of the number of grades in

IRT on the item's functioning with its options. This research contributes to IRT-based studies by analyzing how scales with 4-point ("Strongly Disagree", "Disagree", "Agree", "Strongly Agree"), 5-point ("Strongly Disagree", "Disagree", "Undecided", "Agree", "Strongly Agree"), 6-point ("Strongly Disagree", "Disagree", "Somewhat Disagree", "Somewhat Agree", "Agree", "Strongly Agree"), and 7-point ("Strongly Disagree", "Disagree", "Somewhat Disagree", "Neither Agree nor Disagree", "Somewhat Agree", "Agree", "Strongly Agree") response sets work. The findings indicate that the number of scale points significantly impacts the perceived psychological distance between options, particularly for seven-point scales. In this study, the "Commitment to Profession of Medicine Scale (CPMS)" comprising 4-point, 5-point, 6-point, and 7-point response sets by Aytug Kosan and Toraman (2020), was used. The researchers who developed the CPMS developed this scale with five response categories (strongly disagree, disagree, partially agree, agree, and completely agree). This study examines the psychometric properties of the data obtained from the scale with 4-point, 5-point, 6-point, and 7-point response sets based on IRT.

2. METHOD

2.1. Participants

A total of 2150 medical students from 16 different universities participated in the study. Participants were divided into 4 groups with 560, 544, 502, and 544 medical students. In this study, the CPMS was used as the data collection tool; and the groups were given 4-point, 5-point, 6-point, and 7-point Likert forms of CPMS, respectively. The first group (n=560) was assigned 4-point, the second group (n=544) 5-point, the third group (n=502) 6-point, and the fourth group (n=544) 7-point Likert forms. The distribution of the participants by gender, study year, and university-type variables is given in Table 1.

Table 1. *Descriptive statistics on participants' sex, study year, and university type variables.*

	Variable	4-point	5-point	6-point	7-point
Sex	Female	300	285	294	280
	Male	260	259	208	264
Year	Preparatory	16	---	---	---
	Year 1	114	226	64	104
	Year 2	190	131	53	81
	Year 3	35	54	89	31
	Year 4	28	36	14	69
	Year 5	140	72	188	30
	Year 6	37	25	94	229
University	State	430	402	462	406
	Foundation (Private)	130	142	40	138

2.2. Measurement Tool

The data were obtained using the Commitment to Profession of Medicine Scale (CPMS), which scale was developed by Aytug Kosan and Toraman (2020) and comprised nine items. The original version of the scale has a 5-point Likert structure (strongly agree, agree, partly agree, disagree, strongly disagree). Within the scope of this research, 4-point, 5-point, 6-point, and 7-point forms of the scale were created and applied to four different groups. Aytug Kosan and Toraman (2020) have reported their scale's validity and reliability evidence through exploratory factor analysis (EFA), confirmatory factor analysis (CFA), IRT, Cronbach Alpha, and marginal reliability coefficient. As a result of factor analysis, the structure of the scale was set as 9 items and a single factor.

2.3. Procedure

- The ethics committee approval was obtained for the study.
- This study was approved by the relevant medical faculty rectors and faculty deans.
- The medical faculties to which the CPMS with a 4-point, 5-point, 6-point, and 7-point Likert answer set would be sent was determined.
- Through the faculty deans, the information about the purpose of the research and how the data collection process would be was shared with the students.
- The scales were delivered online to the students who voluntarily agreed to participate and answer the scales.
- The data were taken from the online environment, transferred to statistical software, and analyzed.

2.4. Data Analysis

Data collected from the participants were analyzed on R 4.1.0 (R Core Team, 2021) using mirt 1.35.1 (Chalmers, 2012) and psych 2.1.6 (Revelle, 2021) packages. In addition, MVN 5.9 (Korkmaz et al., 2014) package was used to determine whether the data showed a multivariate normal distribution. In the data analysis, the tested topics, respectively, are:

- Multivariate normality (Henze-Zirkler Test),
- Unidimensionality can be determined by correlation matrix examination or factor analysis while unidimensionality can be determined using factor analytical techniques (Exploratory Factor Analysis [EFA], Principal Axis Factoring [PAF], Eigenvalue),
- The average variance extracted (AVE) and composite reliability (CR) of the scale were investigated for convergent validity. For these two specified values, AVE $\geq$ 0.5 and CR $\geq$ 0.7 are required (Fornell & Larcker, 1981)
- Local independence is a fundamental assumption in item response theory (IRT) models. This assumption states that the responses to one item are independent of the responses to other items at a specific level of ability. This does not imply the absence of correlation between items across all groups; rather it indicates that the responses to an item are independent at different levels of proficiency. To fulfill the local independence assumption, it is essential to meet the one-dimensionality assumption. In a one-dimensional model, if item responses are not locally independent, it indicates a multidimensionality dependency. While one-dimensionality is considered sufficient to meet the local independence assumption, additional methods are employed to specifically assess local independence. One such method is the Q_3 test proposed by Yen (1984). This test evaluates local independence between pairs of items by calculating the residuals of each individual's item responses, based on the estimated item parameters. Yen (1984) recommends that researchers treat items with a linear correlation coefficient exceeding 0.20 as potential violators of local independence. This revised text emphasizes key concepts, uses more precise terminology, and avoids unnecessary repetition. It also integrates the information smoothly and provides a clearer understanding of the concept of local independence in the context of IRT models.
- Item-model fit evaluated with S_χ^2 statistic: According to Browne and Cudeck (1993), the fit indicator in the RMSEA values of the S_χ^2 statistic is considered as 0.05 and below, and according to Hu and Bentler (1999), as 0.06 and below.
- Item-total correlations, internal consistency (Cronbach α), and marginal reliability levels: Hair, et al. (2014), in social sciences, where information is generally less certain, a solution that meets 60% (and sometimes even less) of the total variance is satisfactory. According to Warner (2013), the acceptable limits are between 40% and 70%. While according to Nunnally and Bernstein (1994), sufficient reliability should be at least 0.70 and above.
- Graded Response Model (GRM): GRM is a ranked response model that assumes the same threshold parameters that define the uniform-ordered categorical response formats category

boundaries. The CPMS structure is also suitable for this modeling. For this reason, modeling was done with GRM.

- Item calibrations made with IRT (GRM): GRM is estimated using marginal maximum likelihood (MML); where the scale is fixed using the latent density function g(0) where the mean and variance are constrained. By convention, g(0) is assumed to be the standard normal density (mean zero and standard deviation one) (Smits et al., 2020). In calibration, one aims to train the item parameters in the IRT model using responses from a sample of the target population. Item calibrations were carried out in accordance with the GRM assumption.
- According to Item Response Theory (IRT), the optimal discrimination parameter ("a" parameter) for an ideal scale item should fall between 0.5 and 2. Research suggests that a discrimination parameter within the range of 0.75 to 2.50 is considered acceptable (Flannery et al., 1995).
- The ideal range for item difficulty levels, as represented by the "b" parameter in Item Response Theory (IRT), is typically considered to be between -1.00 and 1.00, indicating a medium difficulty level (Hambleton, 1994). In inability or achievement tests, items with difficulty levels below -1.00 are generally classified as easy, while those with difficulty levels above 1.00 are considered difficult.
- Option Characteristic Curves (OCC) were examined. OCCs correlate the probability of confirming an item's response options with increasing levels of the trait being measured (Sodano et al., 2014).

3. RESULTS

We conducted a multivariate normal distribution test on CPMS datasets containing Likert answer sets with 4-point, 5-point, 6-point, and 7-point scales. The results did not demonstrate multivariate normal distribution. However, factor analysis revealed that the scales exhibit a one-dimensional structure. Table 2 presents the eigenvalues obtained from Exploratory Factor Analysis (EFA), along with the corresponding variance explained, Cronbach's α, AVE, CR, and marginal reliability coefficients.

Table 2. *EFA, explained variance, and reliability coefficients.*

	4-point Likert	5-point Likert	6-point Likert	7-point Likert
KMO	0.784	0.877	0.852	0.848
Bartlett's Test of Sphericity	1583.437 (*df*=36, *p*<.05)	2495.955 (*df*=36, *p*<.05)	2409.577 (*df*=36, *p*<.05)	4804.329 (*df*=36, *p*<.05)
Eigenvalues	3.11	4.46	4.38	6.01
Variance explained	35%	50%	49%	67%
Cronbach α	0.81	0.89	0.89	0.95
r_{jx}	0.85	0.90	0.91	0.95
AVE	0.40	0.49	0.49	0.67
CR	0.85	0.89	0.89	0.95

While the eigenvalue was almost identical in the 5-point and 6-point forms, it increased gradually from the 4-point form to the 7-point form. Similarly, the variance explanation and reliability coefficient increased gradually from the 4-point form to the 7-point form. The variance explained and reliability levels in the 5-point and 6-point forms were very close. There are different opinions about how the factor structure obtained should explain the variance of the desired feature. 5-point, 6-point, and 7-point forms achieved the level of variance explanation suggested by the literature. 4-point, 5-point, 6-point and 7-point forms provided reliability at the level suggested by the literature. AVE and CR rates at the level suggested by the literature occurred in forms with 5-point, 6-point, and 7-point response categories.

Yen's Q_3 statistics (Yen, 1993) were used to determine whether the items met the local independence assumption, and local independence was provided in all four forms. At this stage, 0.20 was used as the criterion value for the Q_3 statistic. The item-model fit was examined with the S_χ^2 statistic. At this stage, the GRM was used as the IRT model. GRM is a polytomous IRT model designed especially for variables accepted as ordinals (Samejima, 2005). The RMSEA values of the S_χ^2 statistic calibrated according to the GRM and showing the item parameters and item model fit are given in Table 3.

Table 3. *Parameter estimation results of CPMS Items.*

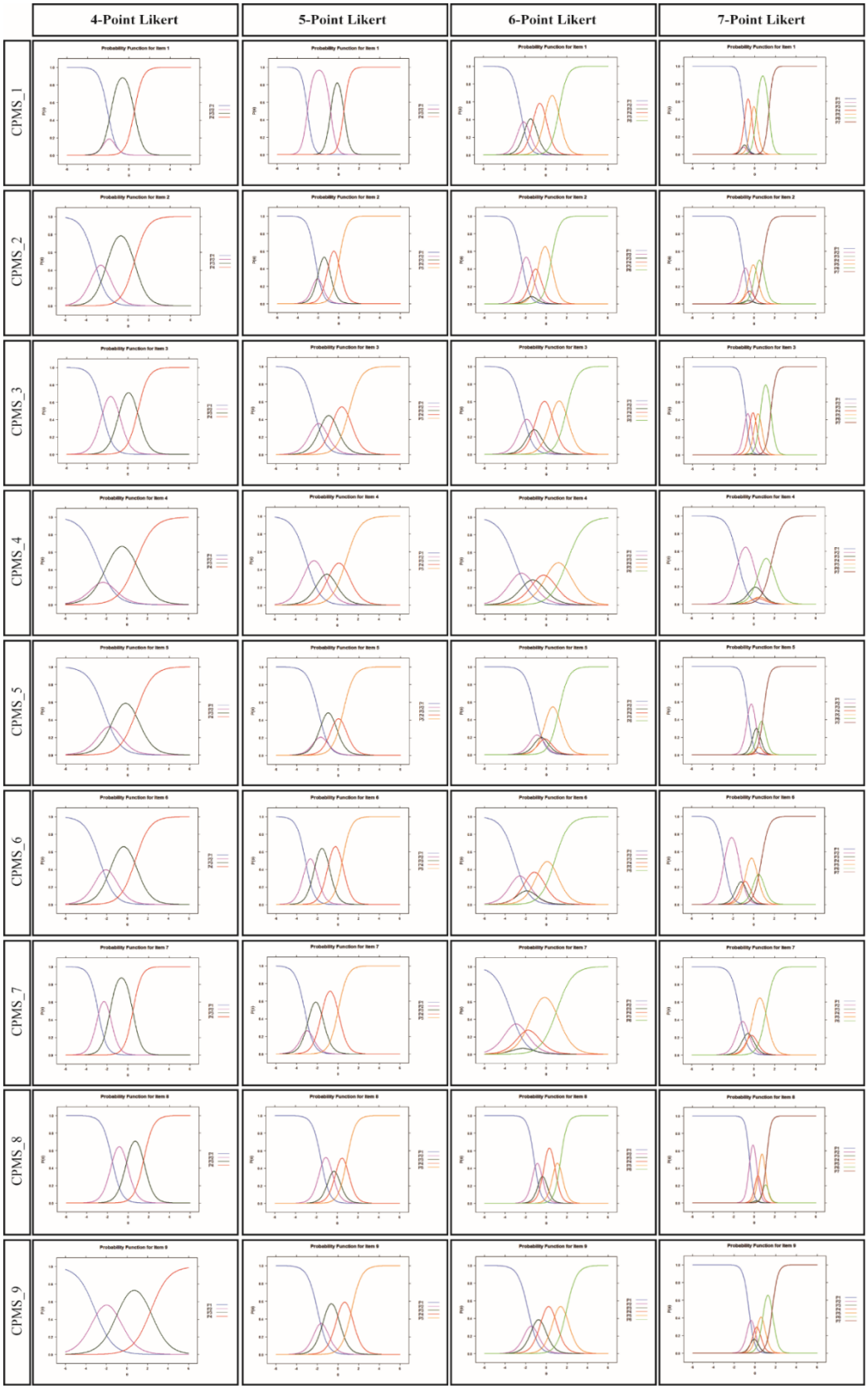
		Items								
		CPMS Item 1	CPMS Item 2	CPMS Item 3	CPMS Item 4	CPMS Item 5	CPMS Item 6	CPMS Item 7	CPMS Item 8	CPMS Item 9
4-point	a	2.45	1.63	1.95	1.16	1.29	1.42	2.45	2.15	1.16
	b_1	-2.06	-3.21	-2.50	-2.79	-2.27	-2.66	-2.87	-1.56	-3.09
	b_2	-1.75	-2.00	-0.85	-1.88	-1.24	-1.48	-1.72	-0.14	-0.89
	b_3	0.53	0.60	0.98	0.90	0.85	0.76	0.49	1.51	2.32
	$RMSEA_{S\_\chi^2}$	0.11	0.12	0.12	0.13	0.11	0.15	0.13	0.11	0.12
5-point	a	3.58	2.74	1.72	1.51	1.96	2.38	2.27	2.47	2.06
	b_1	-3.04	-2.31	-2.31	-2.95	-1.92	-3.16	-3.22	-1.56	-1.98
	b_2	-0.79	-1.45	-1.45	-1.50	-1.49	-2.19	-2.74	-0.62	-1.27
	b_3	0.51	-0.34	-0.34	-0.54	-0.42	-0.92	-1.55	-0.01	-0.01
	b_4	---	1.07	1.07	0.83	0.48	0.42	0.03	0.92	1.32
	$RMSEA_{S\_\chi^2}$	0.09	0.07	0.12	0.07	0.10	0.09	0.11	0.08	0.08
6-point	a	2.50	2.62	1.98	1.23	2.11	1.49	1.32	3.18	2.06
	b_1	-2.43	-2.39	-2.32	-3.03	-1.13	-3.02	-3.47	-1.18	-1.78
	b_2	-1.80	-1.48	-1.47	-1.79	-0.69	-2.11	-2.38	-0.56	-1.15
	b_3	-1.11	-1.35	-0.89	-0.83	-0.31	-1.68	-2.17	-0.17	-0.36
	b_4	-0.04	-0.71	0.52	0.32	0.03	-0.64	-1.32	0.77	0.81
	b5	1.26	0.49	1.95	2.03	1.19	0.80	1.02	1.39	1.96
	$RMSEA_{S\_\chi^2}$	0.13	0.11	0.09	0.13	0.11	0.11	0.10	0.11	0.11
7-point	a	4.81	3.42	4.27	2.05	3.84	2.76	2.57	4.92	3.69
	b_1	-1.09	-1.10	-0.90	-1.54	-0.61	-2.85	-1.40	-0.45	-0.58
	b_2	-1.03	-0.59	-0.42	-0.04	0.08	-1.40	-0.77	0.21	-0.16
	b_3	-0.94	-0.54	-0.41	0.35	0.40	-1.02	-0.38	0.23	0.02
	b_4	-0.32	-0.37	0.08	0.48	0.42	-0.63	-0.03	0.50	0.35
	b_5	0.19	0.20	0.55	0.64	0.51	0.22	1.17	1.02	0.82
	b_6	1.37	0.84	1.56	1.76	0.93	0.74	---	1.19	1.67
	$RMSEA_{S\_\chi^2}$	0.13	0.13	0.17	0.15	0.13	0.12	0.15	0.15	0.15

In the analysis of CPMS data sets applied with 4-point, 5-point, 6-point, and 7-point Likert response sets, the RMSEA values of the S_χ^2 statistic varied between 0.07 and 0.17. The closest fit to the values determined by the literature was obtained in the 5-point Likert form.

There were mathematical differences in the item discrimination "a" parameters of the four forms. It was determined mathematically that the scale items in 4-point and 6-point forms approached the ideal level of discrimination. The increase in the number of grades in the Likert response set of the scale can be said to increase discrimination. In the context of using the Generalized Rating Scale Model (GRM) as an Item Response Theory (IRT) model, the 'b' parameters representing item confirmation difficulty indicate the level of theta at which the likelihood of selecting categories 2 and 3 equals the likelihood of selecting category 1, and the likelihood of selecting category 3 equals the likelihood of selecting categories 1 and 2. The b parameters increased from the first response category to the last response category for all four forms.

Option Characteristic Curves (OCC), item information function, test information function, and reliability functions were obtained after item calibrations. OCCs were examined to better understand how the number of categories changes the response behavior. The OCCs of 4-point, 5-point, 6-point, and 7-point response categories for all items are given in Figure 1.

Figure 1. *OCCs of the items of the CPMS forms administered with a 4-point, 5-point, 6-point, and 7-point Likert answer sets.*



When the Option Characteristic Curves (OCCs) are examined, a summary similar to Table 4 can be made.

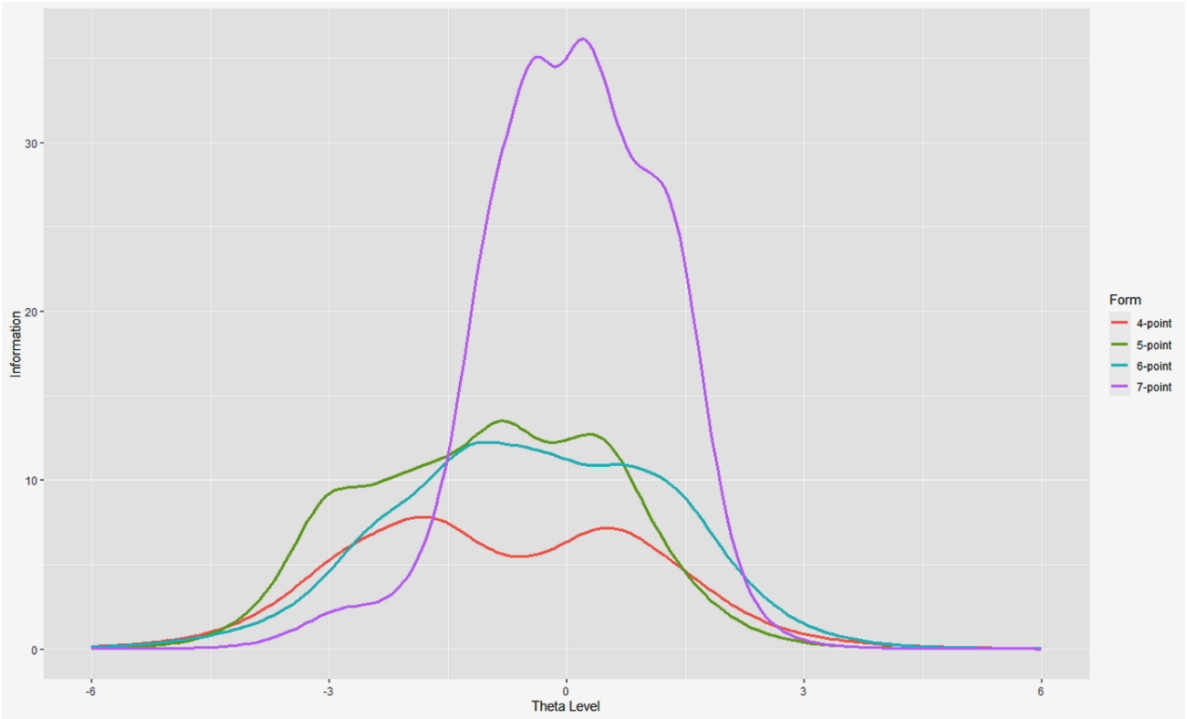
Table 4. *Option functioning states of items according to OCC review.*

Results	4-point Likert	5-point Likert	6-point Likert	7-point Likert
There are Options that Work Well	Items 2, 3, 7, 8, and 9	Items 4, 6, and 8	Items 1, 3, 4, and 8	Items 3, 6, 7, and 9
There is an Option that Never Works	---	Item 1	---	Item 7
There is an Option that Does Not Differ from Other Options	Items 1, 4, and 5	Items 2, 3, 5, 7, and 9	Items 1, 2, 5, 6, 7, and 9	Items 1, 2, 4, 5, and 8
There are very few Responsive Options	Items 1 and 3	Items 2, 5, and 7	Items 2, 5, 6, and 7	Items 1, 2, 4, 5, and 9

The item options differentiated and worked better in the 4-point Likert form. Additionally, in the 5-point and 7-point Likert forms, there was an item with at least one dysfunctioning option. The number of items with an undifferentiated option from other options was the least in the 4-point Likert form. The number of items with options that received a small response from the participating medical school students was also the least in the 4-point Likert form. As seen in Tables 2 and 3, the 4-point Likert form least explained the variance of the scale's measured feature and the item-model fit parameters were not at the level suggested by the literature. However, the 4-point Likert form worked well in identifying the item options and obtaining the participants' responses.

The CPMS forms applied with 4-point, 5-point, 6-point, and 7-point Likert answer sets that gave information with a total of 9 items were examined. The test information functions of the four forms are presented in Figure 2.

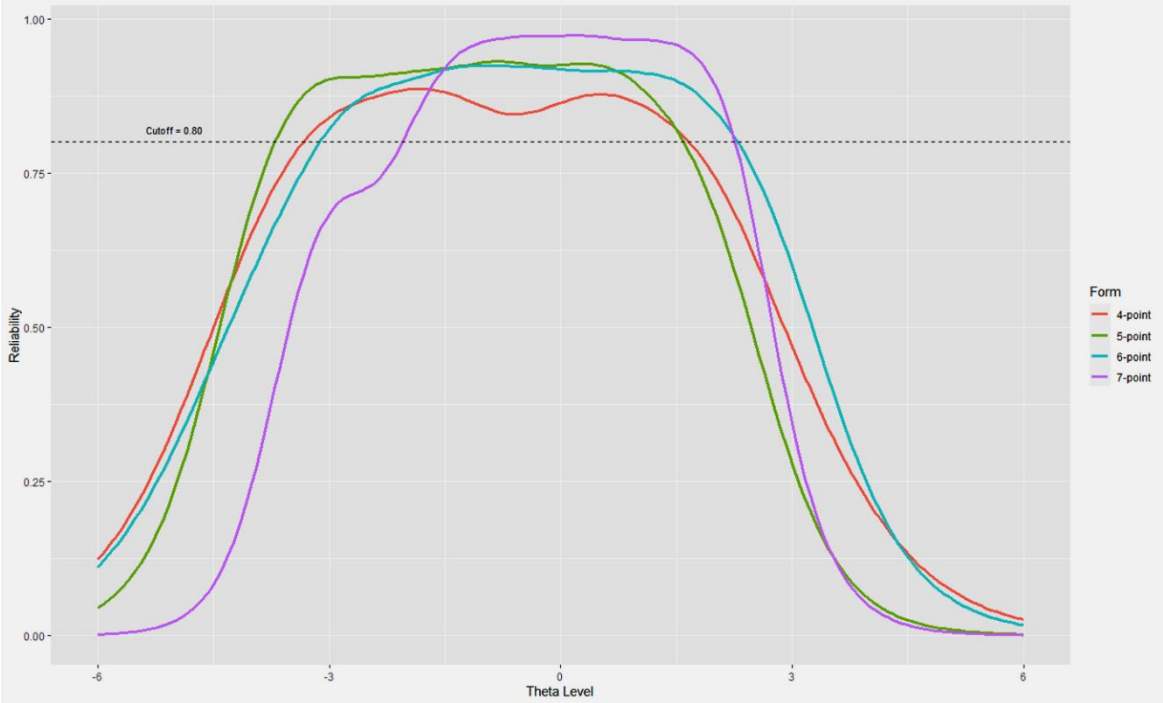
Figure 2. *Test information functions of the four forms.*



When the test information functions are examined, the form that provides the least information is the 4-point Likert form. Additionally, 5-point and 6-point Likert forms gave similar information. However, 5-point and 6-point Likert forms gave higher information than 4-point

Likert forms and lower than 7-point Likert forms. The most informative form was the 7-point Likert form. The reliability functions obtained for the four forms are presented in Figure 3.

Figure 3. *Reliability functions of the four forms.*



When the reliability functions were examined, the levels of all four forms exceeded 0.80 and were reliable at a similar level. The form with the highest reliability was the 7-point Likert form, albeit by a small margin. On the other hand, the 4-point, 5-point, and 6-point Likert forms were similar and had higher internal consistency in a slightly wider theta range compared to the 7-point Likert form.

4. DISCUSSION and CONCLUSION

This study investigates the psychometric properties of data collected using a scale with 4-point, 5-point, 6-point, and 7-point response options, employing Item Response Theory (IRT) as the analytical framework. In the research, data obtained with 4-point, 5-point, 6-point, and 7-point response categories forms were analyzed based on IRT. The psychometric evidence obtained pertains to the information presentation levels of the scale items. While the eigenvalue is almost identical in the 5-point and 6-point graded forms, it increases gradually from the 4-point form to the 7-point form. Similarly, the variance disclosure percentage of the scale's measured feature and the scale data's reliability level have increased gradually from the 4-point to the 7-point form. The variance and reliability levels explained in the 5-point and 6-point forms were very close. In the study by Aybek and Toraman (2022), the reliability coefficient of the scale was calculated for the 4-point, 5-point, and 7-point forms. The more categories a form had, the higher reliability values were reached. In addition, researchers could not obtain a multivariate normal distribution in the data set similar to our study. Leung (2011) applied 4, 5, 6, and 11-point Likert scales in their study and did not find a big difference in Cronbach Alpha value and factor loads. In Chang's (1994) and Preston and Colman's (2000) studies, scales with fewer categories in the response set gave higher reliability values. Prior studies have shown that differences in response categories do not change the Cronbach Alpha coefficient much and that scales with fewer response categories offer a higher level of reliability. In our study, when Figure 3 is examined, it is seen that there is not much difference between the reliability levels. However, as the number of categories decreased, reliability decreased, and as the number of

categories increased, reliability increased. In this respect, it can be said that the study results are compatible with the study conducted by Leung (2011).

The closest fit values to the item-model fits determined in the literature were obtained in the 5-point Likert form. The increase in the number of degrees in the Likert response set in the scale forms increased the discrimination. In this study, the item options differentiated and worked better in the 4-point Likert form. The number of items with the least undifferentiated option is in the 4-point Likert form. The 4-point Likert form had the least items with unspecific responses from medical students. Therefore, the 4-point Likert form explained the variance of the scale's measured feature the least, and the item-model fit parameters were not at the level suggested by the literature. However, the 4-point Likert form performed well in terms of working out the item options and obtaining the participants' responses. In the study by Aybek and Toraman (2022), forms of a measurement tool with 3-point, 5-point, and 7-point response sets were tested. The researchers analyzed the data they obtained based on IRT. The results showed no difference between the three forms in terms of "a" parameters, and the 5-point and 7-point response categories were more advantageous regarding test knowledge and reliability functions. However, seven response categories according to OCCs could not be distinguished by the participants. According to the research of Adelson and McCoach (2010) and Aybek and Toraman (2022), scale forms with 5-point response sets work well. Wakita et al. (2012) applied the forms of a scale with 4, 5, and 7-point response sets to 722 students. The researchers analyzed the data based on IRT. The results showed that the number of degrees of the scale affects the psychological distance between the options, especially for the scale with 7 degrees.

In the present study, an examination of the test information functions showed that the 4-point Likert form provides the least information. The 5-point and 6-point Likert forms gave information close to each other. The 5-point and 6-point Likert forms gave higher information than the 4-point Likert forms and lower than the 7-point Likert forms. The most informative form was the 7-point Likert. When the reliability functions were examined, the reliability level of all four forms exceeded 0.80 and were reliable at a level close to each other. The form with the highest reliability was the 7-point Likert form, albeit by a small margin. On the other hand, the 4-point, 5-point, and 6-point Likert forms were similar and had higher internal consistency in a slightly wider theta range compared to the 7-point Likert form. In the study by Aybek and Toraman (2022), test information and reliability functions showed that using the 7-point response category could provide a better advantage over using the 5-point response.

As a result, increasing the number of degrees in the response sets positively affected the level of informing, and the level of variance explained regarding the feature of interest. However, the 4 and 5-point Likert-type forms were also prominent in terms of better discrimination of options, not less advantageous than the 6 and 7-point forms.

5. LIMITATIONS

In the study, all participants were administered the 4-point, 5-point, 6-point, and 7-point Likert forms of the CPMS at different times (leaving the scale items long enough to be forgotten). This way, data of four different forms could have been obtained from 2150 medical school students. However, the vast majority of the participants did not accept participation in all four different forms. This situation prevented some comparisons (such as comparing the scores of each individual in all forms).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Çanakkale Onsekiz Mart Universtity, Scientific Research Ethics Committee, 03.02.2022 dated 03/11 numbered

Contribution of Authors

Murat Tekin: Investigation, Resources, and Writing-original draft. **Çetin Toraman:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Ayşen Melek Aytuğ Koşan:** Investigation, Formal Analysis, and Writing-original draft.

Orcid

Murat Tekin https://orcid.org/0000-0001-6841-3045
Çetin Toraman https://orcid.org/0000-0001-5319-0731
Ayşen Melek Aytuğ Koşan https://orcid.org/0000-0001-5298-2032

REFERENCES

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

Aiken, L.R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement, 43*, 397-401.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice-Hall International, Inc.

Aybek, E.C., & Toraman, C. (2022). How many response categories are sufficient for Likert type scales? An empirical study based on the Item Response Theory. *International Journal of Assessment Tools in Education, 9*(2), 534-547. https://doi.org/10.21449/ijate.1132931

Aytug Kosan, A.M., & Toraman, C. (2020). Development and application of the commitment to profession of medicine scale using classical test theory and item response theory. *Croatian Medical Journal, 61*(5), 391-400. https://doi.org/10.3325/cmj.2020.61.391

Bora, B. (2013). *A study on the applicability of the likert type scales in marketing*. Doctoral Thesis. Sakarya University. Sakarya.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen K., Long J. (Eds.), *Testing structural equation models* (pp. 136-162). SAGE.

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

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology, 23*, 323-331.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205-215. https://doi.org/10.1177/014662169401800302

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50*(1), 61-104. https://doi.org/10.1177/147078530805000106

DeVellis, R.F. (2003). *Scale development, theory and applications*. SAGE Publications.

Dunn-Rankin, P., Knezek, G.A., Wallace, S., & Zhang, S. (2004). *Scaling methods*. Lawrence Erlbaum Associates, Inc.

Flannery, W.P., Reise, S.P., & Widaman, K.F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnarie II. *Research in Personality, 29*(2), 168-188. https://doi.org/10.1006/jrpe.1995.1010

Fornell, C., & Larcker, D.F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2014). *Multivariate data analysis*. Pearson Education Limited.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological test: A progress report. *European Journal of Psychological Assessment, 10*(3), 229-244.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1217-1218

Joshi, A., Kale, S., Chandel, S., & Pal, D.K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology (BJAST), 7*(4), 396-403. https://doi.org/10.9734/BJAST/2015/14975

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

Leung, S.O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert Scales. *Journal of Social Service Research, 37*, 412-421. https://doi.org/10.1080/01488376.2011.580697

Likert, R. (1932). A technique for the measurement of attitudes. *Arch Psychology, 22*(140), 55.

Lord, F.M. (1954). Chapter II: Scaling. *Review of Educational Research, 24*(5), 375-392. https://doi.org/10.3102/00346543024005375

Mariano, L.T., Phillips, A., Estes, K., & Kilburn, R. (2024). *Should survey Likert Scales include neutral responce categories? Evidence from a randomized school climate survey*. Working Paper. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/working_papers/WRA3100/WRA3135-2/RAND_WRA3135-2.pdf

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Adv in Health Sci Educ 15*, 625-632. https://doi.org/10.1007/s10459-010-9222-y

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. McGraw-Hill, Inc.

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica 104*, 1-15. https://doi.org/10.1016/s0001-6918(99)00050-5

Price, L.R. (2017). *Psychometric methods, theory into practice*. The Guilford Press

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Revelle, W. (2021). *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois. R package version 2.1.6. https://CRAN.R-project.org/package=psych

Samejima, F. (2005). *Graded response model in encyclopedia of social measurement*, edit. Kimberly Kempf-Leonard (pp: 145-153). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00451-5

Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C.B., & Chalmers, R.P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research, 29*(4), 1030-1048. https://doi.org/10.1177/0962280220907625

Sodano, S.M., Tracey, T.J.G., & Hafkenscheid, A. (2014) A brief Dutch language ımpact message ınventory-circumplex (IMI-C Short) using non-parametric item response theory. *Psychotherapy Research, 24*(5), 616-628. https://doi.org/10.1080/10503307.2013.847984

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680

Thomas, H. (1982). IQ interval scales, and normal distributions. *Psychological Bulletin, 91*, 198-202

Torgerson, W.S. (1958). *Theory and methods of scaling*. John Willey & Sons, Inc.

Warner, R.M. (2013). *Applied statistics, from bivariate through multivariate tecniques*. SAGE Publications, Inc.

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*(4), 533–546. https://doi.org/10.1177/0013164411431162

Wong, C.-S., Chuen, K.-C., & Fung, M.-Y. (1993). Differences between odd and even number of response scales: Some empirical evidence. *Chinese Journal of Psychology, 35*, 75-86.

Wu, H., & Leung, S.O. (2017). Can Likert Scales be treated as interval scales? A simulation study. *Journal of Social Service Research, 43*(4), 527-532. https://doi.org/10.1080/01488376.2017.1329775

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3),187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.