International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 534–547
https://doi.org/10.21449/ijate.1132931
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

How many response categories are sufficient for Likert type scales? An empirical study based on the Item Response Theory

Eren Can Aybek[1,*], **Cetin Toraman**[2]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye
[2]Canakkale Onsekiz Mart University, Faculty of Medicine, Department of Medicine Education, Canakkale, Türkiye

ARTICLE HISTORY

Received: Jan. 04, 2022
Revised: Apr. 10, 2022
Accepted: June 18, 2022

Keywords:
Likert-type scale
Response categories,
Item response theory.

Abstract: The current study investigates the optimum number of response categories for the Likert type of scales under the item response theory (IRT). The data was collected from university students attend to mainly the faculty of medicine and the faculty of education. A form of the "Social Gender Equity Scale" developed by Gozutok et al. (2017) was prepared, which had 3, 5 and 7-point response categories. The graded response model (GRM) was used for item calibrations. The results of the study have revealed that using a 5-point response option provides advantages over using a 3-point response category in terms of reliability and test information perspective in the scale development process. The-5 point scale also provides easier responding process for the respondents while it does not pose a major disadvantage compared to a 7-point response category in the terms of reliability. Therefore, based on the findings of the study, researchers are recommended to use a 5-point response category in their scale development process.

1. INTRODUCTION

Using scales is one of the ways to collect data in educational, behavioral and social sciences. Various ways of developing scales have been reported in the literature. The most widely used ones among these include the Thurstone scaling technique (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017; Torgerson, 1958), Guttman scales (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017), and Likert-type rating scales (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Price, 2017).

In the Likert-type item construction, there is a statement related to the psychological trait in concern and a rating showing the levels of agreement with, or approval of, this statement (Anastasi & Urbina, 1997; DeVellis, 2003). Likert scales are widely used in instruments measuring thoughts, beliefs, and attitudes (DeVellis, 2003). Likert's (1932) arguments about Likert-type scales are that a) the distances between categories can be kept equal, b) naming of categories can be arranged beforehand even if they are subjective, and c) the judgements of the prepared scale can be changed according to the item analyses to be carried out on the basis of

*CONTACT: Eren Can AYBEK ✉ erencan@aybek.net Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye

e-ISSN: 2148-7456 /

the responses of those taking the scale (as cited by Dunn-Rankin et al., 2004). To summarize, Likert's arguments are evaluated based on the distribution of real variables (Price, 2017). In Likert scales, response categories are so arranged that their rating distances are equal as much as possible (DeVellis, 2003). A response category may be structured to have 5 ratings in the form of "strongly disagree", "disagree", "indecisive", "agree", and "strongly agree" (Anastasi & Urbina, 1997) as well as 6 ratings in the form of "strongly disagree", "disagree", "somewhat disagree", "somewhat agree", "agree", and "strongly agree" (DeVellis, 2003). There may also be a neutral point among the ratings of a response category. There are proposals for the ratings to be used at this neutral point such as "neither agree nor disagree" or "agree and disagree equally", but debates on how this neutral point should be expressed still continue (DeVellis, 2003). Likert scales are more popular due to the ease of constructing. They are widely used in the social sciences and educational research (Joshi et al., 2015).

In Likert scales, ordinal categorical scores are generated from responses given by the respondents to the scale items. These scores correspond to a basically two-pole range from strongly disagree to strongly agree (Price, 2017). Some researchers argue that the data obtained from a Likert scale are at an ordinal scale level and statistical techniques suitable for such data should be used (Jamieson, 2004; Stevens, 1946; Thomas, 1982). Although an equal intervals assumption is generally made for Likert scales in practice (i.e., distances between the numbers in an ordinal scale), such an assumption often cannot be evidenced from the perspective of measurement essentials/basics. In the face of this dilemma, the question "Should the data be processed on an ordinal scale or an equal interval scale?" is often asked. Norman (2010) pointed out that Likert scales can be accepted at an equal interval scale level and parametric analyses can be used based on this assumption. In their simulation-based study, Wu & Leung (2017) argued that increasing the number of ratings in the response category of a Likert scale would result in a normal distribution and a similarity with an interval scale.

What ratings and denotations should be used when the response category ratings of a Likert-type scale are prepared? How many ratings should be used to exhibit better psychometric features? These and similar questions were the objects of curiosity and major motivations for conducting this study. Studies with similar objects of curiosity have already taken their places in the literature. Aiken (1983) and Wong et al. (1993) have shown that the number of ratings in a response category has no effect on the alpha coefficient. Champney & Marshall (1939) argued that widely used response categories with 5 or 7 points were not appropriate and suggested that points of response categories should be between 18 and 24. In their study, Chang (1994) tried a 9-item scale as a Likert scale with a response category of 4 and 6 points on 165 participants. The purpose of the trial was to compare the reliability values of the scales with a 4-point or 6-point response category. The results of the study showed that the 6-point scale had a decrease in both reliability and heterothrait-monomethod (THMM) correlations. The 4-point scale also had a higher reliability than the 6-point scale in a multitrait-multimethod (MTMM) covariance matrix analysis. In their study, Preston & Colman (2000) gave 149 participants a scale (with ratings between 2 and 11) to evaluate the service of a restaurant they have visited recently. The best psychometric characteristics were exhibited by the scale with a 7-point response category. The test-retest reliability tended to decline in scales with more than 10-point response category. Dawes (2008) investigated how the use of a Likert scale with 5, 7 and 10-point response categories affected the data obtained with respect to arithmetic means and distribution metrics. Three groups of 300, 250 and 185 participants were administered a scale with 5, 7 and 10-point response categories for 8 questions. Each group were given a scale with a different response category. The 10-point format tended to produce lower arithmetic means than the 5 and 7-point formats (the 5 and 7-point formats were converted to be able compare them with the 10-point format). The skewness and kurtosis values of the scales were very close to each other. In a study by Adelson and McCoach (2010), the same mathematics attitude scale with either a 4-point

response category or a 5-point version including a neutral choice was administered to the 3rd and 6th grade students. The study result showed that the 3rd and 6th grade students had the ability to discriminate the 5-point response option. The participants were also found to like the 4-point response option more than the 5-point response option.

Leung (2011) prepared the Rosenberg Self-Esteem Scale in the form of a Likert scale having 4, 5, 6 and 11-point response categories and administered it to 1217 students. A significant difference was not found in the arithmetic means, standard deviations, item correlations, Cronbach Alpha values and factor loadings of the data obtained from these scales of different rating types. The values obtained from the response category with the largest number of ratings (11-point) were found to reduce skewness and kurtosis and produce data close to normal distribution. In the Kolmogorov-Smirnov and Shapiro-Wilk normal distribution tests applied to the study data, 6 and 11-point scales were found to show a normal distribution. In a study conducted by Wakita et.al. (2012), a scale with the same items was administered to 722 undergraduate students in the form of a Likert scale with 4, 5 and 7-point response categories. The analyses in that study were carried out based on the item response theory. The study result showed that the number of points in the scale influenced the psychological distance between the choices, particularly in the 7-point scale. In a study carried out by Bora (2013), the data obtained from the same Likert scale with 5, 7, 9 and 11-point response categories were compared with respect to arithmetic mean, standard deviation, skewness, and kurtosis. The study was conducted with 413 university students. According to the results, increasing number of choices in the response category resulted in decreasing arithmetic means. When the 5-point response category was used, the skewness value was closest to the normal distribution while the kurtosis value was closest to the normal distribution in the 11-point response category.

In summary, the studies in the literature investigated the number of response categories for the Likert type of scales according to reliability, covariance matrices, descriptive statistics, discrimination of neutral category, its effect on factor loadings, and normal distribution based on CTT. Only a study revealed that psychological distance was affected by number of response categories based on IRT. Current study would contribute to the literature by investigating how response categories work under the IRT.

In the present study, a form of the “Social Gender Equity Scale” developed by Gozutok et al. (2017) was prepared, which had 3, 5 and 7-point response categories. The purpose of the study is to investigate the psychometric characteristics of the data obtained from the scale having 3, 5 and 7-point response categories on the basis of the item response theory (IRT).

2. METHOD

2.1. Participants

The participants are students from 11 different universities. The 3-point, 5-point and 7-point Likert forms of the same scale were administered separately group by group to 512, 514 and 498 students, respectively. The number of students who received all of the forms was 153. The distribution of the participants by gender and faculty and their mean ages have been presented in Table 1.

According to Table 1, it is seen that the forms were mostly answered by female students. The students at a medical school and a faculty of education also outnumbered others in each of the three forms. The median age in all three forms was 20.

Table 1. *Descriptive statistics of the participants with respect to their genders, faculties, and ages.*

	3-Point Likert	5-Point Likert	7-Point Likert
Gender	%	%	%
Female	69.53	73.35	73.04
Male	27.54	24.51	25.15
Unknown	2.93	2.14	1.81
Faculty	%	%	%
Medicine	45.90	40.07	38.83
Education	26.17	34.63	29.18
Other	27.93	25.30	31.99
	Median	Median	Median
Age	20	20	20

2.2. Instrument

The data collection tool used in this study was the "Social Gender Equity Scale (SGES)" developed by Gözütok et al. (2017). The scale was administered to two groups of high school students as it was being developed. The first group included 396 high school students. The data obtained from this group were used for the exploratory factor analysis (EFA) and the calculation of Cronbach Alpha reliability coefficient. The second group included 265 high school students, and the data obtained from this group were used for a confirmatory factor analysis. The exploratory factor analysis showed that the scale consisted of 13 items and 2 subfactors, the first of which was "Male Dominance Mentality (MDM)." This factor had 8 items, none of which was reverse scored. The second factor is "Women's Dependence on Men Mentality (WDMM)" which had 5 items and none of them was reverse scored. The 2-factor SGES explains 53% of the total variance about the characteristic in concern (perceived social gender equity). The level of reliability was .882 for the first subscale, .701 for the second subscale and .889 for the whole SGES. The factor construct obtained was validated by a confirmatory factor analysis.

The SGES was used for university students in a study conducted by Toraman and Ozen (2019). A confirmatory factor analysis (CFA) was carried out to see whether the same factor structure explored based on the data obtained from the high school students will be valid for the university students. The result of the CFA confirmed the factor structure of the scale in the university students as well.

2.3. Data Collection

The 3, 5 and 7-point response category forms of the instrument were sent online to the participants within 2-week intervals. Although the primary goal of the data collection process was to have all the participants who could be contacted answer all of the forms, this goal could not be achieved, and the number of participants who received all forms turned out to be 153. However, assuming that all participants had received the three forms, the data collection process was planned in a way to prevent a sequence effect. Accordingly, participants at different universities received the forms in a different sequence. While some participants first took the 3-point form, then the 5-point and finally the 7-point, some others took the 7-point form first, then the 5-point and 3-point forms. In this way, the forms were administered in 6 different sequences. The forms were administered within 2-week intervals. In order to match data from different forms, a nickname, last 4 digits of their phone numbers and last 4 digits of their student numbers were collected from the participants.

2.4. Data Analysis

The data collected from the participants were analysed on R 4.1.0 (R Core Team, 2021) using the mirt 1.35.1 (Chalmers, 2012) and psych 2.1.6 (Revelle, 2021). The MVN 5.9 (Korkmaz, et al., 2014) package was used to see if the data exhibited a multivariate normal distribution. In the analysis of data, first multivariate normality was tested, then unidimensionality was checked using factor analytic techniques. The local independence was tested using Yen's Q3 statistics, and item-model fit was examined based on the S_χ^2 statistics. Finally, item calibration was performed based on the IRT.

A Henze-Zirkler test was performed to test multivariate normality assumption and the data of the three forms were observed not to meet the multivariate normality assumption ($p < .05$). In the exploratory factor analysis (EFA) to test unidimensionality, the Principal Axis Factoring technique was used as the factor exclusion method. When testing unidimensionality, the analysis was limited with a single factor and the Eigenvalues of the first and second factors were evaluated. The item discrimination indices were evaluated using the item-rest correlation and the internal reliability using McDonald's ω. The statistics obtained for each of the three forms are given in Table 2.

Table 2. *EFA results, summary of item statistics and reliability coefficient.*

	3-Point Likert	5-Point Likert	7-Point Likert
Eigenvalues			
First factor	4.046	5.572	5.393
Second factor	.623	.503	.514
Variance explained	31.1%	42.9%	41.5%
McDonald's ω	.852	.906	.900
r_{jx} minimum	.353	.454	.383
r_{jx} maximum	.586	.710	.713

It was seen that the items in all three forms could be combined under a single factor. With a single-factor analysis, 31.1%, 42.9% and 41.5% of the variance in the items of the three forms could be explained. The internal consistency coefficients of the items were over .80, and the item discrimination indices over .30 in all the forms. As such, all three forms were agreed to satisfy the unidimensionality assumption.

Yen's Q3 statistics was used to find out whether or not the items satisfied the local independence assumption, and it was seen that local independence was satisfied in all three forms. At this point, .37 was used as a benchmark for the Q3 statistics. Then, item-model fit was tested based on the S_χ^2 statistics. The RMSEA values of the S_χ^2 statistics ranged between .000 and .063 in all three forms. Thus, it was concluded that the items provided fit to the one factor model in all three forms.

After completing the prerequisite examinations, item calibrations were performed based on the The Graded Response Model (GRM). After calibrating items, item correlations, option characteristic curves (ORF), item information functions, test information function, and reliability functions were obtained.

3. RESULT

In accordance with the aim of the study, the three forms were calibrated based on the GRM. The item parameters and the RMSEA values of the S_χ^2 statistics showing item-model fit are given in Table 3 and Table 4.

A review of Table 3 and 4 reveals that although there are mathematical differences between the three forms in terms of item discrimination (*a*) parameters, the confidence intervals of the *a* parameters in the three different forms are seen to intersect. Therefore, the number of categories in the scale does not change the *a* parameters of the items. Since the GRM was used as an IRT model, item difficulty (*b*) parameters show the theta level that corresponds to the point where the likelihood of choosing category 1 versus 2 and 3, 1 and 2 versus 3 was equal. In all three forms, the *b* parameters showed increase when moving from the first response category to the last response category.

Table 3. *Item parameters for items 1-7.*

		i1	i2	i3	i4	i5	i6	i7
3-Point Likert	*a*	1.462 (.187)	.892 (.139)	1.399 (.184)	2.151 (.305)	1.773 (.241)	2.060 (.368)	2.329 (.380)
	b_1	.617 (.096)	.909 (.161)	.802 (.109)	1.349 (.119)	1.232 (.121)	2.182 (.223)	1.711 (.153)
	b_2	1.845 (.192)	2,308 (.330)	2.361 (.253)	2.319 (.219)	2.062 (.201)	2.738 (.311)	2.389 (.234)
	$RMSEA_{S\_\chi^2}$	.000	.025	.025	.029	.027	.015	.036
5-Point Likert	*a*	1.666 (.149)	1.255 (.124)	1.711 (.155)	2.723 (.248)	2.402 (.212)	3.840 (.451)	4.129 (.463)
	b_1	-.146 (.081)	-.257 (.097)	-.048 (.078)	.346 (.066)	.233 (.068)	.907 (.069)	.870 (.067)
	b_2	.952 (.095)	.963 (.113)	.788 (.088)	1.388 (.095)	1.251 (.092)	1.956 (.132)	1.535 (.093)
	b_3	1.922 (.157)	1.860 (.178)	1.854 (.151)	2.108 (.153)	1.807 (.129)	2.624 (.259)	1.876 (.120)
	b_4	3.576 (.394)	4.214 (.502)	2.842 (.258)	2.861 (.282)	2,562 (.217)	2.926 (.365)	2.603 (.259)
	$RMSEA_{S\_\chi^2}$	.024	.039	.026	.044	.037	.037	.022
7-Point Likert	*a*	1.818 (.161)	1.126 (.120)	1.787 (.163)	2.976 (.282)	2.419 (.218)	2.936 (.346)	4.192 (.483)
	b_1	.018 (.077)	-.142 (.103)	.071 (.077)	.478 (.066)	.274 (.069)	1.092 (.083)	.935 (.069)
	b_2	.843 (.087)	.848 (.117)	.764 (.086)	1.262 (.089)	1.149 (.089)	2.005 (.147)	1.298 (.082)
	b_3	1.082 (.097)	1.163 (.137)	1.060 (.099)	1.414 (.097)	1.320 (.098)	2.078 (.156)	1.443 (.090)
	b_4	1.484 (.120)	1.640 (.175)	1.404 (.120)	1.711 (.117)	1.679 (.122)	2.304 (.188)	1.607 (.101)
	b_5	2.253 (.185)	2.645 (.277)	2.243 (.187)	2.031 (.147)	2.104 (.160)	2.831 (.309)	2.059 (.145)
	b_6	3.630 (.428)	4.286 (.515)	3.364 (.353)	2.710 (.257)	3.418 (.422)	3.062 (.394)	3.460 (.687)
	$RMSEA_{S\_\chi^2}$	.000	.029	.026	.030	.007	.027	.041

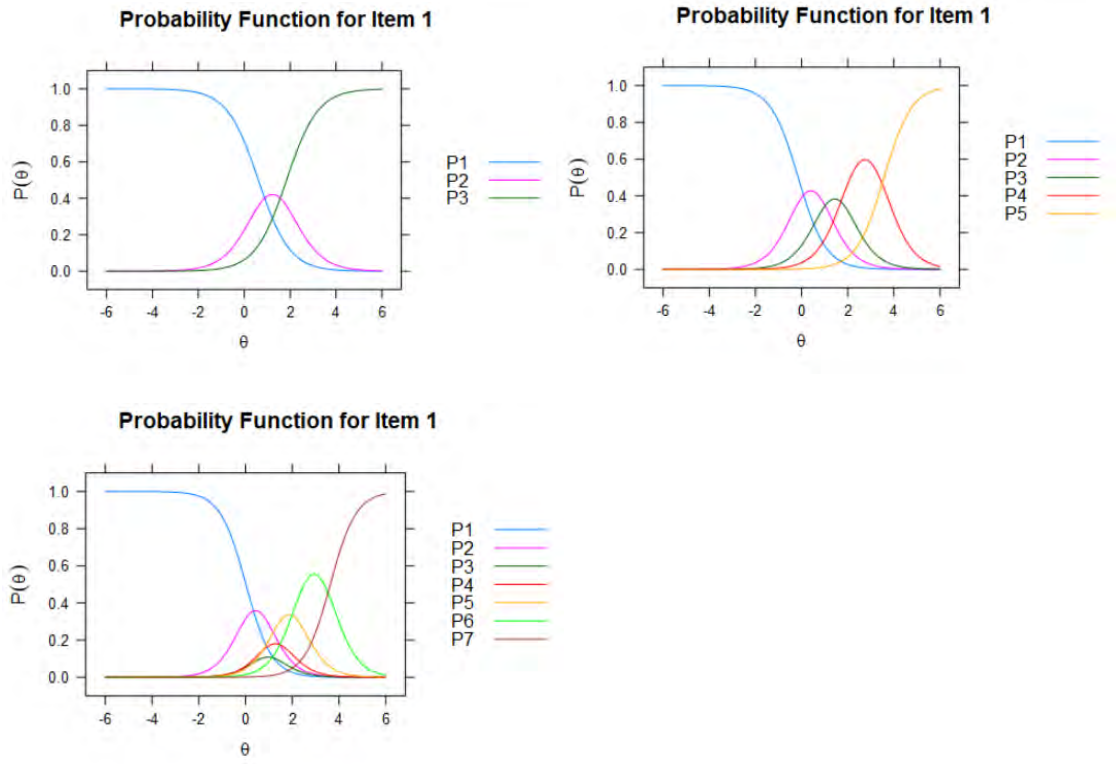
Table 4. *Item parameters for items 8-13.*

		i8	i9	i10	i11	i12	i13
3-Point Likert	*a*	2.112	1.473	3.294	2.352	2.253	1.375
		(.300)	(.215)	(.672)	(.386)	(.364)	(.198)
	b_1	1.335	1.602	2.043	1.862	1.848	1.374
		(.120)	(.174)	(.166)	(.164)	(.165)	(.158)
	b_2	2.374	2.574	2.799	2.728	2.490	2.261
		(.226)	(.292)	(.283)	(.280)	(.250)	(.259)
	$RMSEA_{S\_\chi^2}$	.011	.021	.063	.000	.035	.027
5-Point Likert	*a*	2.484	1.953	3.871	3.832	2.539	2.027
		(.230)	(.187)	(.495)	(.425)	(.274)	(.637)
	b_1	.405	.500	1.136	.870	.970	.637
		(.068)	(.076)	(.076)	(.068)	(.079)	(.077)
	b_2	1.426	1.448	2.036	1.641	1.699	1.509
		(.102)	(.113)	(.145)	(.102)	(.120)	(.116)
	b_3	2.009	2.032	2.670	2.022	2.162	2.066
		(.148)	(.162)	(.276)	(.138)	(.169)	(.164)
	b_4	3.004	3.124	NA	2.944	2.888	3.063
		(.304)	(.318)		(.364)	(.307)	(.310)
	$RMSEA_{S\_\chi^2}$	.014	.017	.040	.043	.040	.013
7-Point Likert	*a*	2.668	2.179	3.514	3.680	2.710	2.140
		(.249)	(.209)	(.450)	(.426)	(.286)	(.214)
	b_1	.489	.558	1.129	.962	.935	.648
		(.068)	(.074)	(.081)	(.072)	(.078)	(.076)
	b_2	1.403	1.288	1.939	1.527	1.672	1.241
		(.098)	(.100)	(.139)	(.099)	(.117)	(.101)
	b_3	1.531	1.398	2.111	1.697	1.743	1.361
		(.106)	(.107)	(.161)	(.112)	(.123)	(.109)
	b_4	1.766	1.637	2.190	1.900	1.952	1.697
		(.123)	(.124)	(.172)	(.133)	(.144)	(.135)
	b_5	2.320	2.129	2.512	2.475	2.513	2.365
		(.183)	(.167)	(.236)	(.226)	(.221)	(.203)
	b_6	3.790	3.015	2.837	2.634	3.774	3.009
		(.612)	(.302)	(.347)	(.264)	(.615)	(.311)
	$RMSEA_{S\_\chi^2}$	.025	.031	.038	.020	.023	.035

The option response functions (ORF) were studied to better understand how the number of categories influenced the response behavior. The ORFs of 3, 5 and 7 response categories for all items are presented in Appendix. The ORFs of the first items of each form are given in Figure 1.

When the ORFs of first items were reviewed, it was seen that each category was differentiated from each other in the forms with 3 and 5 response categories, whereas only 5 of the categories were differentiated in the form with 7 categories. The probability of choosing the third (somewhat disagree) and the fourth (neither agree nor disagree) categories in particular remained lower than the others. This means that when the first item of the scale is presented with 3 and 5 categories, every response category works, whereas when presented with 7 categories, only five categories work. A similar situation can be seen in Appendix.

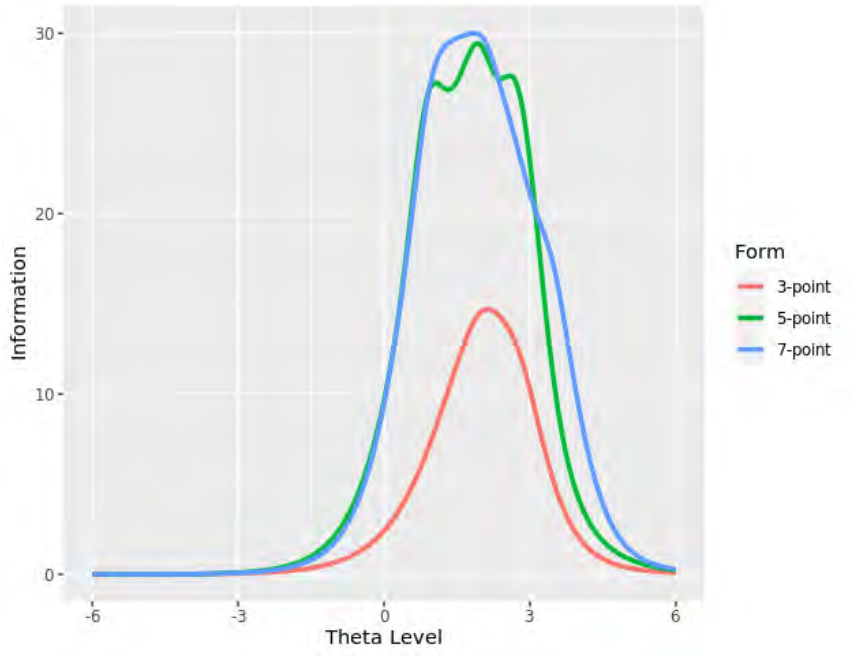
Figure 1. *ORFs of the first items of the forms with 3, 5 and 7 response categories.*



While the middle category (neither agree nor disagree) was not discriminated in items 2, 5, 6, 9, 12 and 13 of the form with 3 response categories, items 2, 5, 6, 7, 9, 11, 12 and 13 in the form with 5 response categories worked as in the 4-category form. The middle category (neither agree nor disagree) was not differentiated from the other categories in the above items except item 6. In item 6, the 'agree' category was not differentiated from the others. As for the 7-response category form, the response categories were not differentiated from each other in all the items. For each item, 4 or 5 categories could be differentiated from each other. This means that the 7-category form was perceived as a 5-category. A similar situation occurred in the test information functions. The test information functions of the three forms are given in Figure 2.

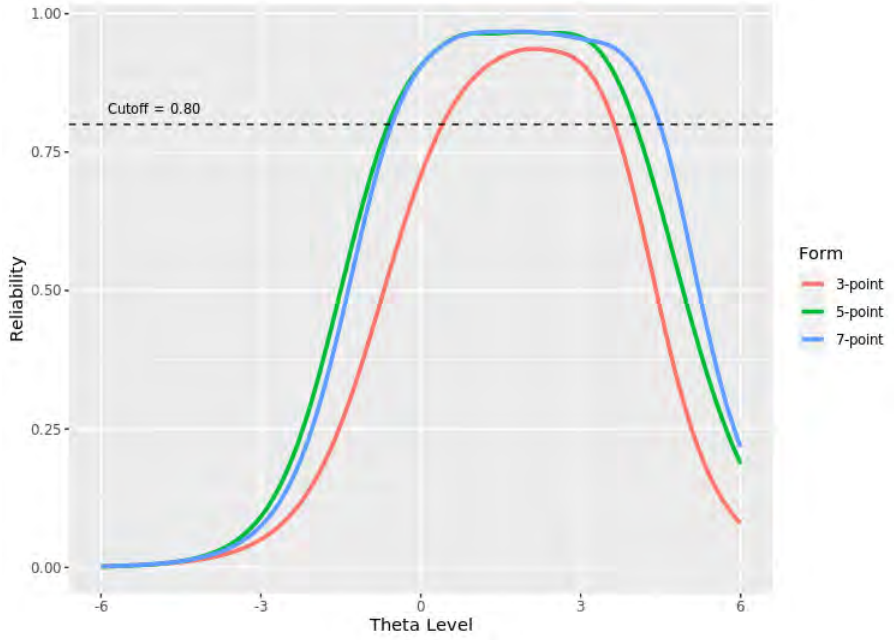
When the test information functions were reviewed, it was understood that the form with 3 response categories provided the least information at a smallest range. The forms with 5 and 7 response categories, on the other hand, provided more information at a much broader range than the form with 3 response categories. The information functions of the forms with 5 and 7 response categories were observed to be quite similar to each other.

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



After reviewing the test information functions, the reliability functions obtained for each of the three forms were also compared and these functions are presented in Figure 3.

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



Supporting the results provided by the test information functions, a review of the reliability functions revealed that the form with 3 categories was able to make measurements with a higher internal consistency at a smaller theta range [.45-3.59]. The forms with 5 and 7 response categories could make measurements with a high internal consistency at a broader range; -.57 to 4.01 for 5-category and -.51 to 4.43 for 7-category. The reliability functions of the forms with 5 and 7 response categories, however, were similar to each other. Nevertheless, these two forms can make reliable measurements for even individuals with fewer peculiarities as compared to the form with 3 response categories.

4. DISCUSSION and CONCLUSION

In conclusion, although there is no difference between the three forms in terms of the *a* parameters, the forms with 5 and 7 response categories are more advantageous in terms of test information and reliability functions. Additionally, the 7 response category could not be differentiated by the participants as shown by the ORFs. The test information and reliability functions showed that using 7 response categories did not provide a significanrt advantage over using 5 response categories.

The number of ratings that is necessary for a response category in Likert-type scales has been debated in the literature. The studies seem to deal with this matter from different points of view. Chang (1994), Preston & Colman (2000) studied it in relation how reliability changed depending on the use of Likert-type scales with response categories having different ratings. Chang (1994) tested a 9-item Likert scale in its 4 and 6-point versions on 165 participants. They found that the 4-point scale had a higher reliability than the 6-point scale. Preston & Colman (2000) obtained the best psychometric outcomes with a scale having a 7-point response category. Studies investigating test-retest reliability have found that test-retest reliability tends to decline in scales with more than 10 ratings. Leung (2011) administered 4, 5, 6 and 11-point versions of a Likert scale to 1217 students. Their study results revealed that there was not a major difference in their Cronbach Alpha values and factor loadings. In the present study, the "McDonald's ω" reliability coefficients of the 3, 5 and 7-point forms of SGES were calculated and the values of .852, .906 and .900 were obtained respectively (see Table 2). Response categories consisting of more categories achieved higher reliability values. In the studies of Chang (1994), Preston & Colman (2000), however, scales with a response category having fewer ratings produced higher reliability values.

Some studies in the literature have focused on what kind of tendency the use of Likert scales with response categories having various ratings created in statistics such as arithmetic mean, normal distribution, skewness and kurtosis. Dawes (2008) investigated in their study how the use of a Likert scale with 5, 7 and 10-point response categories affected the data with respect to arithmetic means and distribution metrics. They found that the 10-point format tended to produce lower arithmetic means than the 5 and 7-point formats (the 5 and 7-point formats were converted to be able compare them with the 10-point format). They obtained very close values between the scales in terms of skewness and kurtosis. Leung (2011) administered a Likert scale with 4, 5, 6 and 11-point response categories to 1217 students. No major difference was found in the arithmetic means, standard deviations, item correlations, Cronbach Alpha values and factor loadings of the data obtained from these scales of different rating types. The values obtained from the response category with the largest number of ratings (11-point) were found to reduce skewness and kurtosis and produce data close to normal distribution. In the Kolmogorov-Smirnov and Shapiro-Wilk normal distribution tests applied to the study data, 6 and 11-point scales were found to show a normal distribution. In a study conducted by Bora (2013), the data obtained from the 5, 7, 9 and 11-point versions of a Likert scale were compared with respect to arithmetic mean, standard deviation, skewness and kurtosis. According to the result of the study, increasing number of choices in the response category resulted in decreasing arithmetic means. As per the skewness value, the scale closest to normal distribution was the 5-point one. As per the kurtosis value, the scale closest to normal distribution was the 11-category scale. In the present study, a multivariate normal distribution could not be obtained in the data obtained from the administration of the 3, 5 and 7-point forms of SGES. In this respect, the results of the present study are not similar to those in the literature.

Some studies in the literature have focused on how participants understand the choices or ratings when Likert scales prepared with response categories having different ratings were used. In a study of Adelson and McCoach (2010), an attitude scale with either a 4-point response

category or a 5-point version including a neutral choice was administered to the 3rd and 6th grade students. The result showed that the 3rd and 6th grade students had the ability to distinguish between the 5-point response option. The participants were also found to favour the 4-point response option more than the 5-point response option. The results of the present study are similar to those of the study carried out by Adelson and McCoach (2010). According to the results, the form where the choices worked best was the 5-point scale form.

Some studies in the literature have focused on how the use of Likert scales prepared with response categories having different ratings affected the choices or ratings based on the Item Response Theory (IRT). In a study conducted by Wakita et.al. (2012), a scale with the same items was administered to 722 undergraduate students in the form of a Likert scale with 4, 5 and 7-point response categories. The analyses in that study were carried out based on the item response theory. The study result showed that the number of ratings of the scale influenced the psychological distance between the choices, particularly in the 7-point scale. As in the study of Wakita et.al. (2012), the present study also used an IRT-based approach. In conclusion, the forms with 5 and 7 response categories are more advantageous in terms of test information and reliability functions than those with 3 response categories. The 7-response category could not be discriminated by the participants.

Although it was originally intended to administer all the forms to exactly the same participants, the number of participants who received all of the three forms remained limited to 153 due to the difficulties in contacting the participants during the pandemic. This may be considered as a limitation of the study, but since the IRT item parameters are group independence, the study was completed as it is. A study where all participants receive all the forms may be designed in further studies. In scale development studies, using a 5-point response option provides advantages over using a 3-point response category, but does not pose a major disadvantage compared to a 7-point response category. Therefore, researchers are recommended to use a 5-point response category, also considering the ease of responding. It is important to conduct similar studies using different scales under IRT so that the generalizability of results can be tested.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship Contribution Statement

Author 1: Finding the problem, literature review, designing the research, data collection, data analysis, and reporting. **Author 2**: Literature review, designing the research, data collection, data analysis, and reporting.

Orcid

Eren Can Aybek https://orcid.org/0000-0003-3040-2337
Cetin Toraman https://orcid.org/0000-0001-5319-0731

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*. The USA: Prentice-Hall International, Inc.

Bora, B. (2013). *Pazarlama araştırmalarında kullanılan likert türü ölçeklerin uygulanabilirliğinin incelenmesi* [*A Study on The Applicability of The Likert Type Scales in Marketing*] [Unpublished doctoral dissertation]. Sakarya University.

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.

Gozutok, F.D., Toraman, C. ve Acar Erdol, T. (2017). Toplumsal cinsiyet eşitliği ölçeğinin (TCEÖ) geliştirilmesi [Development of gender equality scale]. *İlköğretim Online Dergisi (Elementary Education Online), 16*(3), 1036-1048. http://dx.doi.org/10.17051/ilkonline.2017.330240

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

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

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

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*. New York: 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, USA, https://CRAN.R-project.org/package=psych version=2.1.6.

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.

Toraman, C. & Ozen, F. (2019). An investigation of the effectiveness of the gender equality course with a specific focus on faculties of education. *Educational Policy Analysis and Strategic Research, 14*(2), 6-28. https://doi.org/10.29329/epasr.2019.201.1

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

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

APPENDIX

Appendix – 1: Option Response Functions for 3, 5, and 7-point Likert Type Items

Figure 4. *Option response functions for 3-point Likert Type Items*

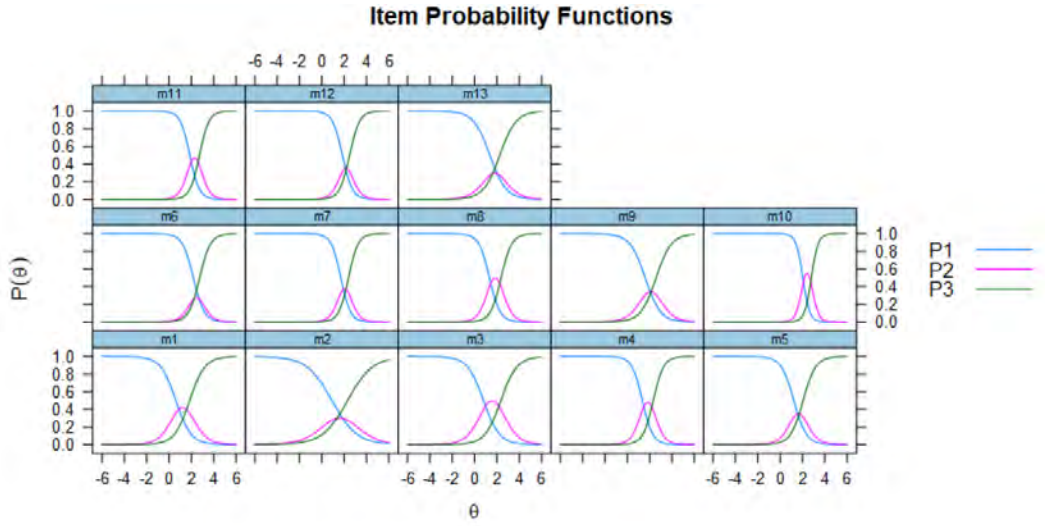


Figure 5. *Option response functions for 5-point Likert Type Items*

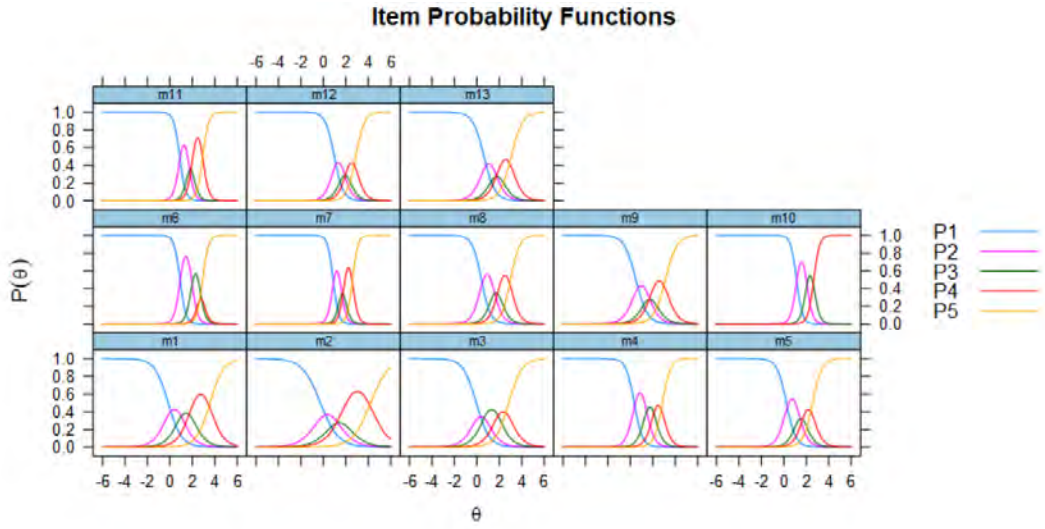


Figure 6. *Option response functions for 7-point Likert Type Items*