International Journal of Contemporary Educational Research (IJCER)

www.ijcer.net

Determining the Factors Affecting the Psychological Distance Between Categories in the Rating Scale

Gizem Uyumaz[1], Gözde Sırgancı[2]
[1]Giresun University,
ORCID ID: 0000-0003-0792-2289
[2]Yozgat Bozok University,
ORCID ID: 0000-0003-4824-5413

Article History
Received: 11.01.2021
Received in revised form: 15.05.2021
Accepted: 06.06.2021
Available online: 03.09.2021

To cite this article:

Uyumaz, G. & Sırgancı, G. (2021). Determining the factors affecting the psychological distance between categories in the rating scale. *International Journal of Contemporary Educational Research, 8*(3), 178-190. DOI: https://doi.org/10.33200/ijcer.858599

International Journal of Contemporary Educational Research
Volume 8, Number 3, September 2021, 178-190 ISSN: 2148-3868

Determining the Factors Affecting the Psychological Distance Between Categories in the Rating Scale

Gizem Uyumaz[1*], Gözde Sırgancı[2]
[1]Giresun University
[2]Yozgat Bozok University

Abstract

In this study, the assumption of the equality of psychological distance between categories of rating scale was tested based on the number of categories and ability distributions. Category parameters were estimated by using generalized partial credit model. The data sets based on the conditions of categories counts and ability distributions were generated by WinGen3 software. The results show that the assumption of the equality of psychological distance between categories of rating scale was not provided in any different ability distribution and different category counts conditions. However, the number of categories influenced the psychological distance between categories, particularly for the 7-point scale. As the number of categories increases, the deviation amount from the conventional category value also increases. Also, endpoints of scales tend to close to middle point of scale when the number of categories is increased. When the converted scale values of the cases with the different ability distribution characteristics were compared, it was seen that the deviation from the conventional category value slightly varied in all the number of categories. However, these differences did not have a systematic order. The degree of violation of the assumption increases as the number of categories increases.

Key words: Interval equality, Number of categories, Ordinal response scale, Likert-type scale

Introduction

Measurement tools used in measuring psychological properties in social sciences are generally ordinal response scale. Scales higher than the ranking scale are almost nonexistent. An ordinal scale consists of values in which order is known, but not the distance between the values. From a mathematical point of view, using ordinal values for calculations such as addition, subtraction, multiplication, and division do not give meaningful results in a strict sense since the distance between two values on an ordinal scale is unknown (Arvidsson, 2019). Because of practical issues or mathematical calculations, many scales used to measure psychological properties are considered equal interval scale, although they are ordinal scale. Therefore, it is assumed that these measurement tools can be scaled and made interval type (Balcı, 2010; Karasar, 2012; Tavşancıl, 2010). Likert scales fall within the ordinal level of measurement (Pett, 1997; Blaikie, 2003; Hansen, 2003). In other words, the response categories are ranked, but the intervals between values cannot be assumed to be equal, although researchers often assume that they are (Blaikie, 2003). However, Cohen, Manion, and Morrison (2000) contend that it is "illegitimate" to assume that the intensity of sentiment between "strongly disagree" and "disagree" is equal to the intensity of sentiment among other consecutive categories on the Likert-type scale. Distance between categories is unequal and unknown (Wu, 2007; Ferrando, 2003; Munshi, 2014), and subjects do perceive Likert-type scales as non-equidistant (Bendixen & Sandler, 1995).

In the social sciences, ordinal variables which have at least five categories and are measured on scales with equal interval length (i.e., equal distances between categories, such as in a Likert-type scale) are regarded as quasi- or pseudo-metric variables, which means they can be treated as metric variables in the empirical analysis (Völkl & Korb, 2018). Pell (2005) suggests that if the Likert-type scale had been conceived to be symmetrical and the assumption of the equal distances between choices had been made, the Likert-type scale may then be expected to provide measures at the interval level. These kind of data is at an ordinal level because the distance

* Corresponding Author: *Gizem Uyumaz, gizemuyumaz@gmail.com*

between categories is not constant throughout the scale; the distance between the 1st and the 2nd categories may not be the same as the distance between the 2nd and the 3rd categories (Harwell & Gatti, 2001). Although a scale needs to be structured in accordance with a coding system based on equal units, it cannot be said with certainty that these numerically equal differences create equal differences in terms of psychological or another basic meaning dimension (Tavşancıl, 2010). Indeed, when using ranking scales, it is often erroneously assumed that equal distances between categories (e.g., the distance between categories 1 and 2 and between categories 2 and 3) correspond to equal distances in measured dimensions (Latorraca, 2018). The main bias in ranking scales is that different respondents associate different meanings to the categories; thus, the perceived distance between categories varies according to the cultural background of the respondents (Crask & Fox, 1987). It is not possible to claim that numerical difference equality (5-4=1)=(4-3=1) exists in terms of semantic difference on a scale where response options are graded from "5.strongly agree" to "1.strongly disagree". Is the difference in meaning between "strongly agree" and "agree" really equal to the meaning difference between "agree" and "neither disagree nor agree"? In the literature, it is argued that for Likert-type scales, only the endpoints labeled scales, the respondents assume that the numerically labeled categories in the middle are equally spaced, and therefore the scale should be seen as interval (Schaeffer & Presser, 2003), fully labeled scales are generally considered ordinal (Wakita, Ueshima & Noguchi, 2012).

Undoubtedly, suppose the unit defined in a scale does not represent the same property in every range of the property being measured. In that case, this means the intervals are not equal, resulting in measurements that cannot accurately represent the quantities of differences across objects/individuals. Thus, the scale's strength to notice the amount of inter-individual differences is gone (Tavşancıl, 2010). In this case, the usability of the results obtained with such a scale and the accuracy of the decisions based on this scale are suspected. Wakita (2004) recommended a method for calculating the distance between categories. The distance between categories (widhts) were defined as W1=C2-C1, W2=C3-C2, W3=C4-C3, and it was displayed that the psychological distances between each category were equal when W1:W2:W3=1:1:1. As a result of the study, it is reported that the distance between categories was affected by the item content. When negative items are used in the scale, the distribution of scores tends to be skewed, and the neutral response category comes near other categories. Another study about the distance between response categories in Likert-type scales with four, five, and seven point scales conducted by Wakita et al. (2012) showed that the equality of distance between response categories differed by number of categories. Thus, they suggest that the differentiation in the number of categories damaged the basic assumption of Likert-type scales. In this case, they stated that the accuracy of the measurements made was at risk. Some studies in the literature show that differentiation of the number of categories affects the validity and the reliability of measurement property (Oskamp, 1977; Tavşancıl, 2010; Tezbaşaran, 1997). The original Likert-type scale contains five verbal response categories. In the rating scales developed later, double, triple, quadruple, six, and seven response categories were also used (Anderson, 1988); however, it is stated that the ideal number of categories is five (Erkuş, 2003; Lozano, García-Cueto & Muñiz, 2008). However, the issue of the optimal number of response categories in rating scales is still unresolved (Preston & Colman, 2000). In fact, one of the main goals in increasing the number of response categories is to obtain high level of internal consistency of the scale (Köklü, 1997). Along with depends on the feature to be measured, to a certain degree, as the number of categories decreases, the sensitivity of the measurement decreases, and it is increases as it increases. After a certain point, either the scale's capacity for distinguishing between categories fails, or information loss occurs because the scale goes to classification level (Erkuş, 2003; 2012). Internal consistency reliability significantly increases until the response category is seven; however, the increase is not remarkable when the number of categories is above seven. Thus, the rating scales, which have over seven categories are mostly not preferred. In other words, when the scale over seven categories is used, it becomes difficult to write a meaningful response category label, and the individual has difficulty in finding a suitable response category (Thorndike, 1997).

In the literature, it is seen that the studies examined psychometric property of scales based on number of response category have different findings. In the studies of Atılgan and Saçkes (2004), Kan (2009), Masters (1974), Matell and Jacoby (1971), Uyumaz and Çokluk (2016), it was found that the differences in the number of categories in Likert-type scales affect the validity of the scale while in the studies of Erkuş, Sanlı, Bağlı and Güven (2000), Leung (2011) and Preston and Colman (2000), it was reported that validity was similar in forms with different category numbers. In the studies of Kan (2009) and Uyumaz and Çokluk (2016), the validity evidence improved as the number of categories increased, while the validity evidence worsened as the number of categories increased in the study of Atılgan and Saçkes (2004). In some studies examining the effect of number of categories on reliability, it was observed that as the number of categories increased, the reliability of the scale increased (Atılgan & Saçkes, 2004; Cicchetti, Showalter & Tyrer, 1985; Masters, 1974; Preston & Colman, 2000; Tate, Simpson, Soo & Lane-Brown, 2011; Uyumaz & Çokluk, 2016; Weng, 2004;), while some of the others showed that reliability did not differ significantly (Bending, 1953; Brown, Wilding & Coulter,

1991; Chang, 1994; Erkuş, et al., 2000; Komorita, 1963; Leung, 2011; Matell & Jacoby, 1971; Uyumaz & Çokluk, 2016; Wakita, et al., 2012).

The usability of the results obtained from any scale, the accuracy and significance of the decisions made are closely related to the psychometric properties of the measurement tool. If a measurement scale does not provide the necessary assumptions and have low validity and reliability, it causes the decisions made about individuals to be inaccurate. In this study, we aimed to examine how the differentiation of the number of categories (four, five, six, and seven) and ability distribution characteristics (normal distribution and beta distribution) affected the distance between response categories in the scale.

Method

Research Model

In this study, how the psychological distances between the categories of rating scale change under different conditions has been explored. In this respect, the research is a simulation study that contributes directly to theoretical studies.

Simulation Design

In order to determine the psychological distances between categories, the Generalized Partial Credit Model (GPCM) (Muraki, 1992) was used. The Partial Credit Model (PCM) is used in measures evaluating the process (problem solving, project evaluation, etc.) and in personality and attitude scales that are scored polytomous. It is seen as an extension of the one-parameter logistic model, and the slopes of the items are considered equal. In GPCM, the slope parameter is estimated separately for each item. The slope parameter describes the degree of differences of categorical response as ability level changes (Embertson & Reise, 2000). GPCM allows more insight into the characteristics of the items than does the PCM. For these reasons, in this study, since the effect of different ability distributions on the psychological distance between categories is examined, GPCM is used for parameter estimation. Item parameters reflecting real-life conditions were chosen. The discrimination parameter "a" was produced from a uniform distribution with a minimum value of 0.5 and a maximum value of 1.5. The difficulty parameter "b" was produced from standard normal distribution with a mean of 0.0 and a standard deviation of 1.0. Similarly, Ogasawara (2001), Paek and Young (2005), Penfield and Bergeron (2005) used a uniform distribution for the "a" parameter in their research. Kim and Lee (2004), Ostini and Nering (2006) determined "b" parameters similarly in this study.

The conditions of the ability distribution are formed by representing the values the situations where the standard normal distribution and the Beta distribution are normal, skewed, and kurtic. The conditions of person parameters are drawn from standard normal distributions which are N(0,1), N(2,1), N(-2,1), N(0,3) and N(0,0.3). These distributions are chosen for representing ability differences of participants and homogeneity and heterogeneity of groups. As suggested by previous IRT literature, the person parameters may be drawn from the same distributions or different distributions (e.g., Dai, 2009, 2013; Li, 2014). In this study, the reason for examining the conditions of the beta distribution in addition to the standard normal distribution is that the Beta distribution is used as conjugate prior for Bernoulli, binomial, negative binomial, and geometric distributions Bayesian inference. The use of the conjugate a priori is quite convenient, as the estimation of the posterior using a conjugate a priori allows avoiding the long and large number of computations of the Bayesian estimation. Thus, we also examined in the effect of different conditions of beta distributions on the psychological distance of rating scales. The conditions reflected different beta distributions are chosen based on previous literature (Moitra, 1990; Pérez, Martín, García & Granero, 2016).

The fixed and manipulated conditions of the study are displayed in Table 1. The normal and beta distribution of the ability distribution in the study, and the conditions in which the number of categories are four, five, six and seven have been examined. Altogether, there were (2x4)x5=40 simulation conditions that were studied. For each condition, 1000 replications were performed, and a total of 40000 data files were examined. The data generation model was GPCM. WinGen3 software was used for data generation (Han, 2007). In the preliminary study, 1000 replications suggested in the litreture for simulation studies are conducted in one condition for the data-generating model. The standard errors of the item and ability parameters were examined for the function of the number of replications. Although the average standard errors show different decreasing or increasing patterns when the replication number ranges till 15, they tend to be stable when the replication number reaches 20. In the study, test lenght and the sample size were kept constant. Moreover, in line with the studies in the literature

(Fitzpatrick & Yen, 2001; Liou, Cheng & Johnson, 1997), all conditions reflect the responses of 1000 participants to 10 items.

Table 1. Fixed and manipulated factors and their corresponding levels in the simulation

Manipulated Factors		
Ability Distribution	Standart Normal Distribution X~N(μ, σ)	Beta Distribution X~Beta(α, β)
	N(0,1): Standard Normal	Be(5,5) : Mesiokurtic
	N(2,1): Standard Normal (High ability)	Be(2,8) : High Positive
	N(-2,1): Standard Normal (Low ability)	Be(8,2) : High Negative
	N(0,3) : Platykurtic	Be(3,3) : Platykurtic
	N(0,0.3): Leptokurtic	Be(8,8) : Leptokurtic
Number of categories	4, 5, 6 and 7	
Fixed Factors		
	Test lenght	10
	Sample Size	1000

Data Analysis

In the data analysis, first of all, McDonald's Omega reliability coefficients were calculated to compare the reliability across conditions. Via a macro for SPSS, IBM SPSS Statistics 26 was used for reliability analysis.

Then, the following steps were applied to determine the psychological distance between response categories:

- Conversion is applied to make conditions with different number of categories comparable. The conversion was made according to the five-point ranking, which is the optimum number of categories. When the item values before the conversion are set at x and those after the conversion are set at y, y=5/4x in the case of the 4-point scale, y=5/6x in the case of the 6-point scale and y=5/7x in the case of the 7-point scale.
- Threshold parameters of all items for each condition were estimated based on GPCM.
- Category parameters were converted to normal distribution, and then standard normal distribution functions were obtained from Equation 1.

$$f(x)=\frac{1}{\sigma.\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \qquad (1)$$

  μ : population mean σ^2 : population standart deviation

- For each number of categories, the scale values were calculated through Equation 2 using the category parameters and category distribution functions of each category.

$$\mu_p=\int_{c_{p-1}}^{c_p} x\frac{f(x)}{\int_{c_{p-1}}^{c_p} f(x)dx}dx=\frac{f(c_{p1-1})-f(c_p)}{\int_{c_{p-1}}^{c_p} f(x)dx} \qquad (2)$$

  c_p : category parameter value of *P*th category, $f(c_p)$: distribution function of *P*th category

- Scale values have been converted into a range of categories through Equation 3.

$$\gamma_n=\frac{p-1}{p_{son}-p_{ilk}}\mu_n \qquad (3)$$

- Converted scale values (d) were obtained by converting the found γ_n values into the category range. If the converted scale values are equal to the category value for all categories, it means that the assumption of the equal interval is provided; in other words, the psychological distances between categories are equal.

Results and Discussion

This section, in parallel with the aim of the study, presents the findings related to the reliability estimates and the determination of the psychological distances between response categories.

Reliability Estimates

The reliability estimates for each condition of normal distribution and beta distribution and for each number of response categories were calculated with McDonald's Omega reliability coefficients, the findings are presented in Table 2.

Table 2. McDonald's omega reliability coefficients

	Normal Distribution					Beta Distribution				
Number of Categories	N(0,1)	N(2,1)	N(-2,1)	N(0,3)	N(0,0.3)	Be(5,5)	Be(2,8)	Be(8,2)	Be(3,3)	Be(8,8)
4	0.999	0.995	0.998	0.999	0.999	0.945	0.966	0.972	0.945	0.955
5	0.998	0.997	0.999	0.999	0.999	0.953	0.972	0.966	0.986	0.911
6	0.998	0.997	0.999	0.998	0.999	0.978	0.967	0.978	0.986	0.923
7	0.998	0.997	0.999	0.998	0.999	0.956	0.963	0.978	0.990	0.656

As shown in Table 2, the reliability estimates are very close to the upper limit of 1.00 and ranges between 0.995 and 0.999 under the normal distribution of the ability distribution. In cases where the ability distribution is compatible with the beta distribution, the reliability coefficients range between 0.656 and 0.990. Independent of the number of categories, reliability in all conditions of the normal distribution is higher than all conditions of the beta distribution.

Different Category Numbers in The Same Ability Distribution

Table 3 to Table 12 shows the threshold parameters (b), the scale value (μ) and the converted scale value (d) obtained by GPCM in different category numbers and ability distribution. If the converted scale value is equal to the category value for all categories, it means that the assumption of the equal interval is provided in practice, in other words, the psychological distances between categories are equal.

The differences between the transformed scale values presented in Table 3 and the conventional item values were calculated for each number of rated response categories. For example, when the number of response categories is four, the amount of deviation from the actual value of the category was calculated by subtracting the converted scale value 2.085 from this value for the conventional category value 2 (| 2 - 2.085 | = 0.085). In this case, the conventional item value of a person who marked the response category "Sometimes" on a graded response scale consisting of four categories such as "Never", "Sometimes", "Often" and "Always" is 2, in fact the response of the individual is 2.085 coincides with the point. Hence, this indicates a difference of 0.085 points between the actual value of the item for that response category and its conventional value. In the N(0,1) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.420, 0.288, 0.009, 0.299, and 0.420, respectively.

Normal Distribution

Table 3. The results of the N(0,1) distribution condition

		Response Categories						
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.854	-0.027	0.821				
	μ	-1.384	-0.373	0.416	1.410			
	d	1.000	2.085	2.933	4.000			
5	b	-1.024	-0.329	0.285	1.052			
	μ	-1.567	-0.637	0.021	0.650	1.544		
	d	1.000	2.196	3.042	3.850	5.000		
6	b	-1.123	-0.472	0.017	0.530	1.179		
	μ	-1.670	-0.825	-0.267	0.223	0.770	1.625	
	d	1.000	2.283	3.129	3.873	4.703	6.000	
7	b	-1.254	-0.633	-0.196	0.198	0.627	1.245	
	μ	-1.724	-0.906	-0.406	-0.001	0.408	0.914	1.732
	d	1.000	2.420	3.288	3.991	4.701	5.580	7.000

(b): Threshold Parameters, (μ): The Scale Value, (d): Converted Scale Value

As the number of categories increased, the amount of deviation from conventional item values increased. Also, it is observed that converted scale values tend to be higher than conventional scale values in low response categories, and converted values tend to be lower than conventional value in high response categories. Another finding is that when the number of response categories are odd, the deviation from the conventional category value is the lowest in the middle category (0.042 in 5 category and 0.009 in 7 category).

Table 4. The results of the N(2,1) distribution condition

		Response Categories						
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.867	0.027	0.872				
	μ	-1.423	-0.423	0.393	1.420			
	d	1.000	2.055	2.917	4.000			
5	b	-1.024	-0.294	0.307	1.012			
	μ	-1.534	-0.633	-0.006	0.630	1.544		
	d	1.000	2.172	2.985	3.812	5.000		
6	b	-1.152	-0.480	-0.005	0.485	1.141		
	μ	-1.639	-0.784	-0.235	0.238	0.786	1.648	
	d	1.000	2.301	3.136	3.855	4.689	6.000	
7	b	-1.260	-0.622	-0.195	0.210	0.629	1.273	
	μ	-1.748	-0.919	-0.413	-0.007	0.402	0.910	1.737
	d	1.000	2.427	3.298	3.997	4.702	5.576	7.000

In the N(2,1) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.427, 0.298, 0.003, 0.298, and 0.424, respectively.

Table 5. The results of the N(-2,1) distribution condition

	Response Categories							
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.833	-0.011	0.840				
	μ	-1.399	-0.390	0.399	1.393			
	d	1.000	2.084	2.932	4.000			
5	b	-1.024	-0.324	0.283	1.073			
	μ	-1.584	-0.643	0.020	0.647	1.544		
	d	1.000	2.203	3.051	3.853	5.000		
6	b	-1.164	-0.484	-0.008	0.508	1.190		
	μ	-1.679	-0.817	-0.245	0.241	0.793	1.658	
	d	1.000	2.293	3.150	3.877	4.704	6.000	
7	b	-1.273	-0.628	-0.186	0.211	0.632	1.221	
	μ	-1.705	-0.900	-0.415	-0.013	0.400	0.918	1.748
	d	1.000	2.399	3.241	3.941	4.658	5.558	7.000

In the N(-2,1) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.399, 0.241, 0.059, 0.342, and 0.442, respectively.

Table 6. The results of the N(0,3) distribution condition

Response Categories								
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.793	0.037	0.860				
	μ	-1.414	-0.424	0.357	1.362			

	d	1.000	2.070	2.914	4.000			
5	b	-1.024	-0.311	0.292	1.006			
	µ	-1.530	-0.622	0.009	0.640	1.544		
	d	1.000	2.182	3.003	3.823	5.000		
6	b	-1.156	-0.493	-0.005	0.503	1.172		
	µ	-1.665	-0.807	-0.244	0.244	0.795	1.651	
	d	1.000	2.293	3.143	3.878	4.709	6.000	
7	b	-1.238	-0.609	-0.162	0.242	0.659	1.266	
	µ	-1.742	-0.933	-0.444	-0.040	0.379	0.893	1.719
	d	1.000	2.402	3.250	3.951	4.677	5.569	7.000

In the N(0,3) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.402, 0.250, 0.049, 0.323, and 0.431, respectively.

Table 7. The results of the N(0,0.03) distribution condition

	Response Categories							
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.845	-0.022	0.804				
	µ	-1.371	-0.369	0.410	1.402			
	d	1.000	2.084	2.926	4.000			
5	b	-1.024	-0.299	0.302	1.026			
	µ	-1.546	-0.636	-0.001	0.633	1.544		
	d	1.000	2.179	2.999	3.820	5.000		
6	b	-1.175	-0.493	-0.001	0.519	1.175		
	µ	-1.667	-0.817	-0.253	0.242	0.802	1.667	
	d	1.000	2.274	3.120	3.863	4.703	6.000	
7	b	-1.314	-0.647	-0.220	0.182	0.631	1.256	
	µ	-1.734	-0.913	-0.400	0.019	0.427	0.945	1.782
	d	1.000	2.400	3.276	3.990	4.687	5.571	7.000

In the N(0,0.3) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.400, 0.276, 0.010, 0.313, and 0.429, respectively.

Beta Distribution

Table 8. The results of the Be(5,5) distribution condition

	Response Categories							
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.876	-0.038	0.821				
	µ	-1.384	-0.368	0.431	1.426			
	d	1.000	2.085	2.937	4.000			
5	b	-1.024	-0.293	0.297	1.006			
	µ	-1.529	-0.624	-0.002	0.629	1.544		
	d	1.000	2.178	2.988	3.810	5.000		
6	b	-1.133	-0.495	-0.002	0.504	1.187		
	µ	-1.677	-0.813	-0.246	0.243	0.787	1.633	
	d	1.000	2.304	3.162	3.901	4.722	6.000	
7	b	-1.299	-0.659	-0.220	0.205	0.622	1.255	
	µ	-1.733	-0.908	-0.408	0.007	0.433	0.946	1.769
	d	1.000	2.413	3.270	3.981	4.710	5.590	7.000

In the Be(5,5) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.413, 0.270, 0.019, 0.290, and 0.410, respectively.

Table 9. The results of the Be(3,3) distribution condition

	Response Categories							
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.871	-0.008	0.853				
	μ	-1.409	-0.397	0.413	1.423			
	d	1.000	2.072	2.930	4.000			
5	b	-1.024	-0.306	0.297	1.042			
	μ	-1.559	-0.639	0.004	0.637	1.544		
	d	1.000	2.186	3.015	3.830	5.000		
6	b	-1.195	-0.517	-0.012	0.498	1.164		
	μ	-1.658	-0.800	-0.237	0.259	0.824	1.683	
	d	1.000	2.283	3.126	3.869	4.714	6.000	
7	b	-1.289	-0.657	-0.217	0.192	0.608	1.271	
	μ	-1.745	-0.905	-0.394	0.013	0.429	0.941	1.761
	d	1.000	2.437	3.312	4.008	4.722	5.597	7.000

In the Be(3,3) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.437, 0.312, 0.008, 0.278, and 0.403, respectively.

Table 10. The results of the Be(8,8) distribution condition

	Response Categories							
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.893	-0.004	0.859				
	μ	-1.414	-0.402	0.419	1.440			
	d	1.000	2.063	2.927	4.000			
5	b	-1.024	-0.286	0.324	1.059			
	μ	-1.573	-0.661	-0.019	0.625	1.544		
	d	1.000	2.170	2.994	3.821	5.000		
6	b	-1.137	-0.504	0.035	0.531	1.179		
	μ	-1.671	-0.826	-0.277	0.229	0.793	1.636	
	d	1.000	2.277	3.107	3.873	4.726	6.000	
7	b	-1.289	-0.642	-0.199	0.196	0.624	1.258	
	μ	-1.735	-0.910	-0.404	0.002	0.414	0.932	1.761
	d	1.000	2.416	3.285	3.981	4.688	5.578	7.000

In the Be(8,8) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.416, 0.285, 0.019, 0.312, and 0.422, respectively.

Table 11. The results of the Be(2,8) distribution condition

	Response Categories							
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.831	0.003	0.825				

	μ	-1.387	-0.391	0.390	1.391			
	d	1.000	2.075	2.919	4.000			
5	b	-1.021	-0.283	0.306	1.048			
	μ	-1.564	-0.647	-0.011	0.625	1.544		
	d	1.000	2.180	2.998	3.816	5.000		
6	b	-1.179	-0.518	-0.009	0.491	1.144		
	μ	-1.641	-0.789	-0.236	0.258	0.818	1.670	
	d	1.000	2.287	3.122	3.868	4.714	6.000	
7	b	-1.242	-0.657	-0.205	0.211	0.656	1.240	
	μ	-1.721	-0.921	-0.426	-0.003	0.424	0.923	1.723
	d	1.000	2.393	3.256	3.993	4.737	5.607	7.000

In the Be(2,8) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.393, 0.256, 0.007, 0.263, and 0.393, respectively.

Table 12. The results of the Be(8,2) distribution condition

Response Categories								
Number of Categories		1.	2.	3.	4.	5.	6.	7.
4	b	-0.875	0.007	0.842				
	μ	-1.400	-0.401	0.406	1.426			
	d	1.000	2.061	2.918	4.000			
5	b	-1.023	-0.287	0.314	1.042			
	μ	-1.559	-0.649	-0.013	0.626	1.544		
	d	1.000	2.173	2.992	3.817	5.000		
6	b	-1.147	-0.452	0.037	0.522	1.178		
	μ	-1.669	-0.820	-0.274	0.203	0.768	1.644	
	d	1.000	2.282	3.105	3.825	4.678	6.000	
7	b	-1.246	-0.644	-0.192	0.203	0.633	1.244	
	μ	-1.724	-0.909	-0.411	-0.005	0.411	0.917	1.725
	d	1.000	2.416	3.282	3.989	4.714	5.594	7.000

In the Be(8,2) distribution condition, the largest distance between the conventional values of the response categories and the transformed scale values happened when the number of response categories was seven. The deviations from the second, third, fourth, fifth and sixth categories were 0.416, 0.282, 0.011, 0.286, and 0.406, respectively.

Different Ability Distributions in The Same Category Numbers

The findings on how the psychological distances between response categories vary according to the ability levels of individuals and the number of response categories of the scale are evaluated together in four figures below.

Figure 1. Converted scale value of the 4-point scale for the normal and beta distribution.

In Figure 1, it is seen that the deviation from the conventional values of the categories in all conditions of both beta and normal distribution is very close to each other and at most 0.085.

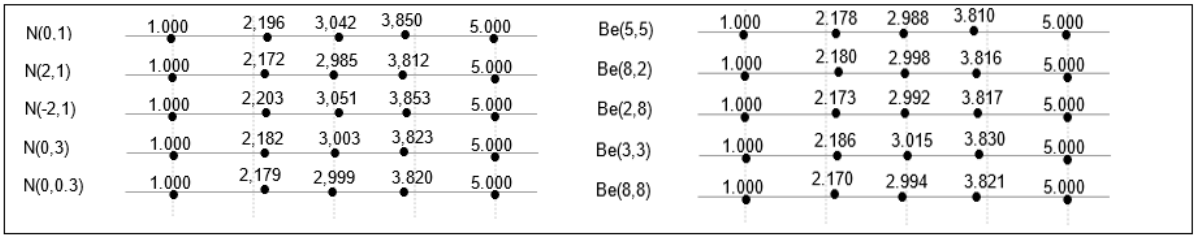

Figure 2. Converted scale value of the 5-point scale for the normal and beta distribution.

Figure 2 summarized that the deviations from the conventional values of the categories are close to each other in all conditions of both normal and beta distributions. The highest devision in the normal distribution is 0.203, and the highest devision in the beta distribution is 0.19. The lowest deviations in both distributions were in the middle category. Usually the endpoint categories tend to close towards the mean.

Figure 3. Converted scale value of the 6-point scale for the normal and beta distribution.

Figure 3 summarized that the deviations from the conventional values of the categories are close to each other in all conditions of both normal and beta distributions. The highest devision in the normal distribution is 0.311, and the highest devision in the beta distribution is 0.322. The highest deviations in both distributions were seen in the 2nd and 5th categories while the lowest deviation was seen in the 3rd and 4th categories. Usually the endpoint categories tend to close towards the mean.

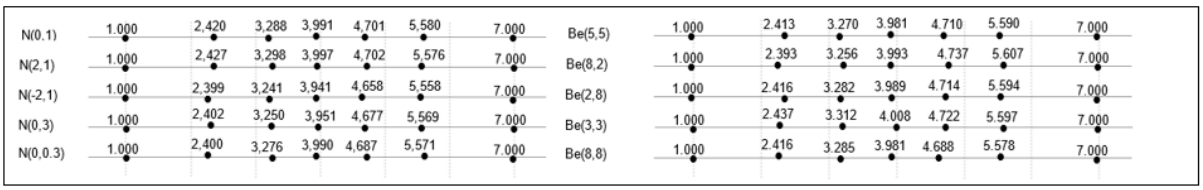

Figure 4. Converted scale value of the 7-point scale for the normal and beta distribution.

Figure 4 summarized that the deviations from the conventional values of the categories are close to each other in all conditions of both normal and beta distributions. The highest devision in the normal distribution is 0.442, and the highest devision in the beta distribution is 0.422. In both distributions, the lowest deviation was seen in the middle category, and the endpoint categories tend to be close to middle. In the beta distribution, the highest deviation from the conventional value of the response categories was observed in the leptokurtic distribution of beta. Particularly, it is seen that the most deviasion from conventional category parameters displayed in the 7-point scale in all conditions of both normal and beta distribution. This deviasion was from the endpoint catogories to midpoint category. Wakita, et al. (2012) illisturated that in the 7-point scale, the participants tended to avoided selecting both ends of categories. These findings show that unequal distance between categories on a 7-point scale may happen respondents' biases against categories containing the strongest statements, not because of the ability distribution of respondents.

Conclusion

Likert-type scales, which are frequently used in social sciences to measure emotional characteristics, have an important assumption. It is the equality of psychological distances between categories. In this study, this assumption was examined by calculating the interval values. It is examined how the number of rating response categories in the scale and the differentiation of ability distribution characteristics of the group affected the reliability of the measurement tool and the psychological distances between the response categories in the scale.

First of all, it has been concluded that the reliability coefficient is at a high level in all types of the normal distribution, independent of the number of categories. In conditions where the ability distribution is compatible with the normal distribution, the reliability values do not change in all conditions of both the distribution and the number of categories. Therefore, in cases where the ability distribution is compatible with the normal

distribution, it has been determined that the reliability coefficient is independent of the number of categories. This result is supported by the studies in the literature (Brown, Wilding & Coulter, 1991; Chang, 1994; Komorita, 1963; Wakita, Ueshima & Noguchi, 2012). When the ability distribution is the beta distribution, reliability coefficient value differs according to the shape of the distribution and the number of categories. However, these differences do not show a consistent structure. Also, the internal consistency reliability coefficient was calculated lower than the normal distribution in every condition of the beta distribution. The beta distribution is skewed compared to the normal distribution. Therefore, it is concluded that the reliability decrases independent of the number of categories when the distribution is skewed. In conditions where the ability distribution is compatible with the beta distribution, the reliability values are affected by the number of categories; in other words, it differs as the number of categories changes. However, this change is not consistent (increasing/decreasing) in the conditions. Hamby and Levine (2016) found that the reliability estimates were higher in conditions where the number of categories was higher than four. Preston and Colman (2000) found that 4-point scale performed poorly on reliability in comparison to scales with more levels. Although there are studies in the literature stating that as the number of categories in the rating scales increases, the reliability of the scale increases (Atılgan & Saçkes, 2004; Cicchetti, Showalter & Tyrer, 1985; Masters, 1974; Preston & Colman, 2000; Tate, Simpson, Soo & Lane-Brown, 2011; Uyumaz & Çokluk, 2016; Weng, 2004), this study shows that the equality of the psychological distances between categories breaks as increases the number of the categories in a scale. For this reason, taking the purpose of the study and the level of the group into account, it seems beneficial not to choose response options containing too many categories.

In sum, it is seen that as the number of response categories increases, the amount of deviation from the conventional category value increases in Likert-type scales. In all cases, it was found that the assumption of the equality of psychological distance between categories could not be fully provided, and there were deviations from the conventional category value in each category. However, the conditions in which the number of categories is fewer are closer to fulfilling the assumption. As a result of this study, it was found that the differentiation in the number of categories in the rating response scales disturbs the equality of psychological distance between categories, which is the basic assumption of the scale. The degree of violation of the assumption increases as the number of categories increases. Therefore, the accuracy of the findings and decisions taken from the applied scale at risk. This finding is in line with the study of Hamby and Levine (2016); Wakita et al. (2012).

In this study, it was determined that the deviations from the category values towards the midpoint category at all category numbers, in other words, conventional category values came over towards the midpoint. While the deviation from the category value is higher in the endpoint categories, it decreases as it gets closer to the midpoint category. In the case of an odd number of categories (five and seven), the deviation from the category value is low in the midpoint category. At the same time, it is high in even number of categories. Thus, one must find an accomondation between potentially increasing the imprecision associated with rating scales by using too many categories and inviting the error of extreme responses by using fewer categories. For this purpose, using 5-point end scales is recommended Hamby and Levine (2016).

Wakita (2004) examined psyhological distance in real data study and reported that when the distribution is skewed, the neutral response category came over towords to other categories (the distance between them decreases). This study determined that in conditions where the ability distribution is skewed to the right and left, the midpoint response category is almost equal to its category value (or the deviation is very small). Also, it was found that the deviations in the right and left categories of the midpoint category change to approach the midpoint category.

When the converted scale values of the situations with the same number of categories and different ability distribution characteristics were examined, it was seen that the deviation from the category value slightly varied in all the number of categories according to the characteristics of the distribution, However, these differences did not have a systematic order.

This study aimed to examine whether the number of options and different ability distributions had an effect on the psychological distance between categories in the Likert scale by applying IRT theory to consider the appropriate number of options. The results of IRT analysis indicated that when the number of options increased, pyshological distance between categories differed from conventional category value, especially in the 7-point scale. Also, it was found that ability distribution did not effect the pyshological distance between categories. However, this study was conducted as simulation study. These results should be supported with real data examinations.

Acknowledgements or Notes

A part of this study was presented 5th Congress of Measurement and Evaluation in Education and Psychology, Antalya, Turkey.

References

Anderson, L. W. (1988). *Likert scales; educational research, methodology, and measurement: An international handbook.* Pergamon. Edited by: Keeves, J. P. (Ed), 227-228

Arvidsson, R. (2019). On the use of ordinal scoring scales in social life cycle assessment. *The International Journal of Life Cycle Assessment, 24*(3), 604-606.

Atılgan, H., & Saçkes, M. (2004). Ölçeklerin ikili ve çok kategorili puanlanmasının psikometrik özelliklerinin karşılaştırılması [Comparison of psychometric properties of dual and multi-category scoring of scales]. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 5*(7).

Balcı, A. (2010). *Sosyal bilimlerde araştırma yöntem, teknik ve ilkeler [Research methods, techniques and principles in social sciences]*. Ankara: PegemA.

Bending, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and the number of categories on the scale. *Journal of Applied Psychology, 37*, 38-41.

Bendixen, M., & Sandler, M. (1995). Converting verbal scales to interval scales using

Blaikie, N. (2003). *Analyzing quantitative data.* London: SAGE Publications Ltd., London.

Brown, G., Wilding, R. E., & Coulter, R. L. (1991). Customer evaluation of retail salespeople using the SOCO scale: A replication extension and application. *Journal of the Academy of Marketing Science*, *9*, 347-351.

Chang, L. (1994). A psychometric evaluation of four-point and six-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*, 205-215.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: A Monte-Carlo investigation. *Applied Psychological Measurement,* 9, 31-36.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education*. 5th edn. London: RoutledgeFalmer.

Crask, M. R., & Fox, R. J. (1987). An exploration of the interval properties of three commonly used marketing research studies: a magnitude estimation approach, *Journal of the Marketing Research Society, 29*(3), 317-39.

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

Erkuş, A. (2003). *Psikometri üzerine yazılar [Writings on psychometry]*. Türk Psikologlar Derneği.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme-1, temel kavramlar ve işlemler [Measurement and scale development in psychology-1, basic concepts and operations]*. Ankara: PegemA.

Erkuş, A., Sanlı, N., Bağlı, M. T., & Güven, K. (2000). Öğretmenliğe ilişkin tutum ölçeği geliştirilmesi [Developing an attitude scale toward teaching as a profession]. *Eğitim ve Bilim*, *25*(116), 27-32.

Ferrando, P. J. (2003). A Kernel density analysis of continuous typical-response scales. *Educational and Psychological Measurement, 63*, 809-824.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. Applied Measurement in Education, 14 (1), 31-57.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement,* 31 (5), 457-459.

Hansen, J. P. (2003). CAN'T MISS-Conquer any number task by making ımportant statistics simple. Part 1. Types of variables, mean, median, variance, and standard deviation. *J. Healthcare Qual, 25*(4), 19-24.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131.

Kan, A. (2009). Effect of scale response format on psychometric properties in teaching self-efficacy. *Eurasian Journal of Educational Research, 34*, 215-228.

Karasar, N. (2012). *Bilimsel araştırma yöntemi: kavramlar, ilkeler, teknikler [Scientific research method: concepts, principles, techniques]*. Nobel Yayın Dağıtım.

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004-5). Iowa City, IA: Act, Inc.

Komorita, S. S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology, 61*, 327-334.

Köklü, N. (1997). Tutumların ölçülmesi ve Likert tipi ölçeklerde kullanılan seçenekler [Measuring attitudes and options used in Likert-type scales]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 28*(2).

Latorraca, R. (2018). Think aloud as a tool for implementing observational learning in the translation class, *Perspectives*, *26*(5), 708-724, DOI: 10.1080/0907676X.2017.1407804

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*, 37:4, 412-421.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the Kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*(4), 349-369, DOI: 10.1177/01466216970214005.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4(2), 73-79. http://dx.doi.org/10.1027/1614-2241.4.2.73

Masters, E. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires 1. *Journal of Educational Measurement, 11*(1), 49-53.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Moitra, S. D. (1990). Skewness and the beta distribution. *Journal of the Operational Research Society*, *41*(10), 953-961.

Munshi, J. (2014). A method for constructing Likert scales, *Social Science Research Network*. doi:10.2139/ ssrn.2419366.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Pell, G. (2005). Use and misuse of Likert scales. *Medical Education, 39*(9), 970. https://doi.org/10.1111/j.1365-2929.2005.02237.x

Pérez, J. G., Martín, M. D. M. L., García, C. G., & Granero, M. Á. S. (2016). Project management under uncertainty beyond beta: The generalized bicubic distribution. *Operations Research Perspectives, 3*, 67-76

Pett, M. A. (1997). *Nonparametric statistics for health care research*. London: SAGE Publications.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologia, 104*, 1-15.

Tate, R. L., Simpson, G. K., Soo, C. A., & Lane-Brown, A.T. (2011). Participation after acquired brain injury: clinical and psychometric considerations of the sydney psychosocial reintegration scale (SPRS). *Journal of Rehabilitation Medicine, 43*(7), 609–618.

Tavşancıl, E. (2010). *Tutumların ölçülmesi ve SPSS ile veri analizi [Measuring attitudes and data analysis with SPSS]* (4. baskı). Nobel.

Tezbaşaran, A. (1997). *Likert tipi ölçek geliştirme kılavuzu [Likert type scale development guide]*. Türk Psikologlar Derneği.

Thorndike, R. (1997). *Measurement and evaluation in psychology and education*, Prentice-Hall.

Ogasawara, H. (2001). Standart errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25*(1), 53 67.

Oskamp, S. (1977). *Attitudes and opinions.* Prentice-Hall.

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

Paek, I., & Young, M. J. (2005). Investigation of student growth recovery in a fixed item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education, 18*(2), 199-215.

Penfield, R. D., & Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement, 29* (3), 218-233.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual review of sociology, 29*, 65-88 https://doi.org/10.1146/annurev.soc.29.110702.110112

Uyumaz, G., & Çokluk, Ö. (2016). An Investigation of Item Order and Rating Differences in Likert-Type Scales in Terms of Psychometric Properties and Attitudes of Respondents. *Journal of Theoretical Educational Science*, *9*(3), 400-425. DOI: 10.5578/keg.10011

Völkl, K., & Korb, C. (2018). Deskriptive Statistik [descriptive statistics]. Wiesbaden: Springer.

Wakita, T. (2004). The distance between categories in rating-scale method: Applying item response model to the assessment process. *Japanese Journal of Psychology, 75*, 331-338.

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.

Weng, L-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement,* 64(6), 956-972.

Wu, C-H. (2007). An empirical study on the transformation of Likert scale data to numerical scores. *Applied Mathematical Sciences*, *1*(58), 2851-2862.