Ankara University Journal of Faculty of Educational Sciences
Year: 2024, Volume: 57, Issue:1, 251-288
DOI: 10.30964/auebfd.1280498, E-ISSN: 2458-8342, P-ISSN: 1301-3718

Examining the Effect of Differently Labeled Likert-Type Scales on Measurement Invariance[1]

ARTICLE TYPE	Received Date	Accepted Date	Published Date
Research Article	04.10.2023	07.16.2023	03.06.2024

Nuri Doğan [2]
Hacettepe University

Ceylan Gündeğer Kılcı [3]
Aksaray University

Meltem Yurtçu [4]
Inonu University

Abstract

This study examined the psychometric properties of three verbal forms with varying midpoint labels and one with end anchored. For this purpose, data were collected from 377 university students using four different scale forms measuring the same feature, which consisted of the same items with different attitude labels. This study used a 14-item short form of the Mathematical Attitude Scale, which has a unidimensional structure. The data were tested for validity, reliability, and measurement invariance. The results indicated that the explained variance rates of the four forms were close and that the forms had the same and high level of reliability. According to the Confirmatory Factor Analysis results, Form 1 with the mid-point label "I have no idea" was in better compliance with the data. Finally, the forms only provided structural invariance. The lack of metric invariance shows that item factor loadings differ across forms. Based on this result, it can be said that individuals perceive different attitude labels in the same category and different types of scales in different ways. Within the scope of these findings, researchers can repeat similar studies based on generalizability theory and/or item response theory to put empirical evidence in the field.

Keywords: Likert, label, undecided, no idea, neither agree nor disagree, end anchored, scale.

Citation: Doğan, N., Gündeğer Kılcı, C., & Yurtçu, M. (2024). Examining the effect of differently labeled Likert-type scales on measurement invariance. *Ankara University Journal of Faculty of Educational Sciences, 57*(1), 251-288. https://doi.org/10.30964/auebfd.1280498

---

[1]This study was presented as an "oral presentation" at the 7th International Congress on Measurement and Evaluation in Education and Psychology (CMEEP-2020) held in Turkey (online congress) between 1- 4 September, 2021.

[2]Prof. Dr., Hacettepe University Education Faculty, Department of Measurement and Evaluation in Education, E-mail: nuridogan2004@gmail.com, https://orcid.org/0000-0001-6274-2016

[3]*Corresponding Author*: Assistant Prof. Dr., Aksaray University Education Faculty, Department of Measurement and Evaluation in Education, E-mail: cgundeger@gmail.com, https://orcid.org/0000-0003-3572-1708

[4]Assoc. Prof. Dr., Inonu University Education Faculty, Department of Measurement and Evaluation in Education, E-mail: meltem.yurtcu@gmail.com, https://orcid.org/0000-0003-3303-5093

Summated rating is a frequently used technique for measuring psychological characteristics such as personality and attitude. This approach is commonly referred to as the Likert method (Anastasi, 1982). In Likert-type scales, the respondent reports the degree of agreement or disagreement with the attitude item covered by each statement in the scale (Tezbaşaran, 2008). Seashore and Katz (1982) stated that the Likert technique is seemingly easy but quite complex in practice. This technique enables the preparation and response of attitude statements. The complexity of the technique lies in obtaining a one-dimensional scale, exploring the dimensionality of attitudes, and using item correlation, factorial, and multidimensional analysis procedures to evaluate the structural and causal relationships between the measured variables (Seashore & Katz, 1982). The formats of attitude labels, which are frequently used in Likert-type scales, are divided into three categories: *verbal*, *end-anchored*, and *numerical*. Verbal scales involve labeling all scale points with a verbal statement. The end-anchored scales' first and last scale points have a verbal label, and the rest are numerical. In numerical scales, all scale points have a numerical value, and sometimes, these scale points can be named with specific percentage labels (0% of the time, 25% of the time, etc.) (Newstead & Arnold, 1989).

In general, there are five categories ranging from *Strongly Agree* (5) to *Strongly Disagree* (1) on Likert-type scales (Jamieson, 2004). Although the verbal scales of five-point Likert-type scales are listed as *Strongly disagree, Disagree, Undecided, Agree, and Strongly agree*, there is not much information in the literature about how this order and labels ought to be. For instance, Robie et al. (2022) stated that measurement invariance, scale mean, and respondent response do not change much when categories are ordered differently. Similarly, although some studies offering the option "I do not know" in terms of representing the respondent's lack of knowledge (Payne, 1950; Vaillancourt, 1973), it is not possible to say that this option strengthens reliability (Krosnick & Presser, 2010). Therefore, researchers decide the order of scale categories and labels based on their knowledge and experience during the scale development stage. In addition to the order of options and labels, the expression "I am undecided" is mainly used in studies as the mid-point of the scale (Kağıtçıbaşı, 2010). The mid-point is labeled as "Neither agree nor disagree" in some studies (Gegez, 2010; Kurtuluş, 2006; Yükselen, 2003), whereas it is labeled as "I have no idea" in others (Nakip, 2006). It is a requirement that the labels at the scale points be clear and understandable for the scale to be highly reliable. If there are ambiguous statements on the labels, the validity and reliability of the scale may be compromised (Krosnick & Presser, 2010).

There is no consensus in the literature regarding whether the degrees of attitude labels are similar, whether there are differences between them in terms of language (semantics), or whether these expressions should be understood as synonymous by individuals or not. From the past to the present, various debates have come to the fore, especially about what the mid-point indicates. According to Başar (2021), expressions such as "I am undecided, I have no idea, I cannot say anything" indicate a situation, not a middle level or trend; therefore, it may be wrong to use such expressions on

scale applications. Instead of these expressions, Başar (2021) states that labels such as "I am impartial", "I am neutral", or "I agree at a moderate level" should be used to indicate agreement or disagreement. However, according to the literature, Likert (1967) and Bogardus (1967) excluded the word "Undecided" in the midpoint in their later studies, and Thurstone (1967) and Allen & Kenney (1967) used the word "Neutral" for the middle option (as cited in Başar, 2021, p.3). Tezbaşaran (2008) and Turgut and Baykul (1992) also stated that they used the word "Undecided" to mean "I am in the middle, I am impartial, I am neutral, I have no idea". According to Başar (2021), "Neutral" carries the meaning of impartial and indicates a decision: "I have made my decision: I am not on one side, I am in the middle". However, "I am undecided" does not indicate a decision. It shows that "I do not have a decision" when the researcher asks what the decision is. Those who do not make a decision cannot be assigned a score indicating the degree of decision (Başar, 2021). Bora Semiz and Altunışık (2016) stated that "Undecided", and "Neither agree nor disagree" labels can be used for the mid-point, but the statement "I have no idea" does not indicate a place in the attitude for the mid-point of the scale, so it would not be appropriate for the midpoint. Moreover, another study has shown that the respondents might also turn to the center, to the midpoint option used in the sense of "Neutral", due to the bias toward the center and the desire to be liked socially (Nadler et al., 2015).

Differences in scale labels can result in differences in the total score obtained from the scales, the reliability and validity of the scales, and the variability in scale responses (Newstead & Arnold, 1989). According to the study conducted by Dixon et al. (1984), the scores obtained from the verbal and end-anchored Likert-type scales did not show a significant difference according to the labels. Similarly, in their studies comparing different label formats of Likert-type scales, Finn (1972) and Wyatt & Meyers (1987) found that label definitions did not reveal a significant difference in scale scores and that the reliability of scales with different labels was similar. On the other hand, Jacko and Huck (1974), in contrast to these studies, found that end-anchored scales had lower mean scores than verbal and numerical scales. Similarly, Krosnick and Berent (1993) and Weng (2004) stated that the reliability of end-anchored scales were lower than that of verbal scales. Considering the studies conducted on numerical scales, Blumberg et al. (1966) revealed no difference between end-anchored and numerical scales. Peters and McCormick (1966) showed that verbal scales had higher reliability than numerical scales (as cited in Newstead & Arnold, 1989, p.35). On the other hand, Newstead and Arnold (1989) stated that numerical scales had a higher mean score than end-anchored scales, but there was no significant difference in mean scores between verbal and end-anchored scales.

In the literature, Likert-type scales have been compared in terms of reliability, primarily based on the number of categories in the scales. In some studies, reliability increases as the number of categories increases (Alwin, 1992; Bandalos & Enders, 1996; Hartley & MacLean, 2006; Kılıç, Uysal, & Kalkan, 2021; Lee et al., 2002; Simms et al., 2019). On the other hand, another study has shown that the reliability does not show a significant difference depending on the number of response

categories and that the five categories give the optimal reliability (Finn, 1972). In addition, in some studies, higher reliability has been obtained with the forced-choice scale (Bendig, 1954), while some scholars have stated that adding a midpoint to the scale would strengthen reliability (O'Muircheartaigh et al., 2000). In summary, a limited number of studies have examined the validity and reliability of end-anchored scales and verbal scales with different attitude labels, and these studies have not yielded consistent results. The level of measurement invariance of scales with the same purpose, consisting of the same scale items but with different labels, has not been extensively studied in the literature.

Measurement invariance means that groups at the same level in the measured feature have the same raw score from the measurement tool or that individuals in different groups perceive and interpret the scale items similarly (Bryne & Watkins, 2003). According to Gregorich (2006), four hierarchical invariance types are mentioned in measurement invariance. The first is structural invariance. In structural invariance, the equality of the scale's factor structure in groups is tested. Metric (weak) invariance can be tested on the condition that structural invariance is ensured. In metric invariance, whether or not the groups perceive the scale items similarly is checked. Failure to provide metric invariance indicates that item factor loads are not equal in the groups. Therefore, if metric invariance cannot be achieved, comparison of the scores obtained from the groups may be biased. In the case where metric invariance is achieved, scalar (strong) invariance is tested. Scalar invariance examines whether item factor loads and regression constants are equal between the groups or not. In the case where scalar invariance is reached, strict invariance is considered, and equality of error variances in groups is tested. The four types of measurement invariance listed are treated as prerequisites for each other. In this case, the failure in structural invariance means that other measurement invariances are not achieved. Therefore, comparing the scores obtained from a measurement tool in which structural invariance cannot be achieved is biased.

The current study aimed to examine the validity and reliability evidences of the scale forms (verbal and end anchored five-point Likert-type scales) consisting of the same items with different category labels with the data obtained from the same student group, to compare these evidences over the forms, and to what extent the forms provide measurement invariance. Considering the limited number of research studies on this topic in the literature, this study can be regarded as important and essential in presenting empirical evidence for the different uses of category labels in scales. This research sought answers to the following sub-problems:

1. What are the item factor loads and Cronbach's alpha reliability coefficients of scales with different category labels as calculated from the results of exploratory factor analysis (EFA)?
2. What is the difference between the item factor loads calculated from the EFA results of scales with different category labels?

3. What is the relationship between the item factor loads calculated from the EFA results of scales with different category labels?
4. How do the total scores obtained from scales with different category labels differ?
5. How do the factor scores obtained from scales with different category labels differ?
6. What is the difference between the fit indices of scales with different category labels as calculated from the results of Confirmatory Factor Analysis?
7. To what extent do scales with different category labels provide measurement invariance?

Method

In this part of the research, the research model, study group, data collection tools, ethics committee decision, data collection process, and data analysis are included.

Research Model

This research is a descriptive study that examines the validity and reliability of scale forms consisting of the same items and having different category labels. Descriptive studies attempt to answer the question "What is…?" (Balcı, 2011). In addition, the research is correlational in terms of revealing the relationship between item factor loads calculated from different forms. Correlational studies attempt to reveal the degree of the relationship with the correlation coefficient (Balcı, 2011).

Study Group

The study group consists of 377 individuals studying at the undergraduate level in different universities in Türkiye in the 2020–2021 academic year. The demographic information about the study group is given in Table 1, below. According to Table 1, most of the study group involved students from Education Faculties Furthermore, most students were studying in the second and third year of the university, and the majority were females.

Table 1
Demographic Information of the Study Group

Characteristic	Category	n	%
Gender	Female	272	72.1
	Male	105	27.9
	Total	377	100
Faculty	Education Faculty	346	91.8
	Sports Faculty	25	6.6
	Other	6	1.6
	Total	377	100
Year	1st Year	57	15.1
	2nd Year	138	36.6
	3rd Year	161	42.7
	4th Year	21	5.6
	Total	377	100

Data Collection Tools

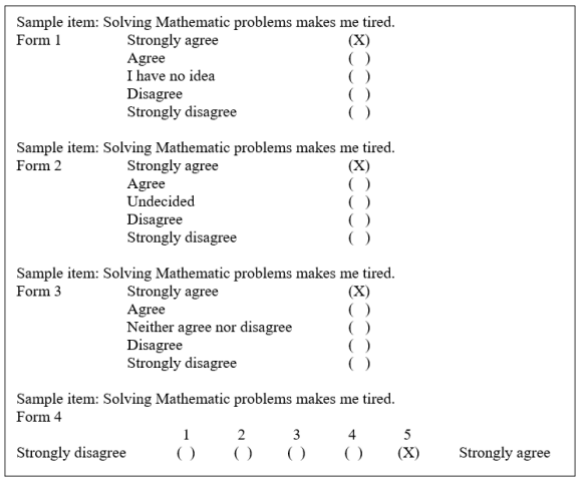
The Mathematical Attitudes Scale (MAS) was used as the data collection tool in the study. MAS, developed by Baykul (1990), is a one-dimensional scale involving 30 items, 15 of which are positive and 15 are negative. The variance rate explained by one dimension of MAS is 56%, and its reliability coefficient is .96. The lowest score obtained from the MAS is 30, and the highest score is 150 (as cited in Nartgün, 2002, p.47).

Within the scope of this research, the scale was shortened because it would take time to apply the four different forms of this 30-item scale to students, and some items in the scale were not found to be sufficient or were considered to be very specific. Since the aim of this study was not to make inferences about the scale scores of the students, the validity and reliability evidences were collected within the scope of the research, considering the effect of shortening the scale on the construct validity. In shortening the scale, attention was paid to removing the expressions that directly point to the thoughts about the Mathematics lesson (*for example, Mathematics is among my favorite subjects*), while the items obtained with a high factor load in Nartgün (2002) remain in the scale. Two more interesting items, believed to measure attitudes toward mathematics, were added to the shortened scale. These items are "*Expressing a situation mathematically makes me happy*." and "*Mathematical discoveries fascinate me*.". The number of items on the scale applied to the students was 14. Six of them represent negative thoughts toward mathematics, while eight of them represent positive thoughts. Four different scale versions were formed by differentiating the only the category labels. The difference in the category labels of these versions, called Form 1, Form 2, Form 3 (which are verbal forms), and Form 4 (end anchored form), is presented in Figure 1, below.

As seen in Figure 1, only the label showing the midpoint was changed in the first three verbal forms. Form 1 involved the expression "I have no idea", Form 2 involved

the expression "I am undecided", and Form 3 involved the expression "Neither agree nor disagree" for the midpoint. In addition, Form 4 used an end-anchored scale, whose beginning and end values are shown with verbal labels, and the remainder consists of numerical values.

Figure 1
Used Forms



Ethical Committee Approval

Before the data collection process, the necessary permits were obtained from the Inonu University Scientific Research and Ethics Committee (Protocol No: 13-19, Date: 02/07/2021). In addition, participation in the research was based on volunteerism by requesting verbal consent.

Data Collection Process

The data were collected online at approximately 1-week intervals. The links of the scale forms were shared with the students, and then the link was closed to them at the end of the allowed time (four days) during the data collection process. Thus, the students were prevented from viewing the forms earlier or later. The students were required to write the first six digits of their identity cards to ensure matching the forms. At the end of the data collection process, the dataset was controlled, and repeated data (filling out a form more than once) were removed from the dataset.

Data Analysis

Before data analysis, a total of 40 individuals detecting univariate and/or multivariate outliers were excluded from the dataset while testing factor analysis assumptions. The descriptive statistics of 337 individuals who answered all forms without repetition are given in Table 2. According to the skewness and kurtosis values in Table 2, the responses did not deviate from the normal distribution. The mean and median values of the variables are close to each other. The minimum and maximum values obtained from the forms indicate that the score range covers all attitude levels. The standard deviation values are similar in the first three forms (verbal forms), and the value slightly increases in the fourth (end anchored) form. Therefore, according to Table 2, the forms provide similar information.

Table 2
Descriptive Statistics of the Forms

Statistics	Form 1	Form 2	Form 3	Form 4
N	337	337	337	337
Mean	45.98	45.89	45.63	46.08
Median	48.00	47.00	47.00	48.00
Mod	56.00	55.00	55.00	66.00
Standard Deviation	14.61	14.47	14.62	15.80
Skewness	-0.33	-0.30	-0.31	-0.30
Standard Error	0.13	0.13	0.13	0.13
Kurtosis	-0.91	-0.87	-0.85	-1.02
Standard Error	0.27	0.27	0.27	0.27
Minimum	14.00	14.00	14.00	14.00
Maximum	70.00	70.00	70.00	70.00

For the first sub-problem, EFA based on the polychoric correlation matrix was applied to the data set, and the reliability of the forms was determined by calculating the Cronbach's alpha internal consistency coefficient due to the unidimensionality of the scale. For the second sub-problem, the difference between the item factor loads obtained from EFA was tested using Friedman and Wilcoxon tests, and the effect sizes of the tests were calculated. For the third sub-problem, the relationship between item factor loads was examined using the Spearman Correlation Coefficient. The differences between the total scores and factor scores obtained from the forms were tested with one-way ANOVA for the fourth and fifth sub-problems.

Confirmatory factor analysis (CFA) and multi-group CFA, a frequently used method to test measurement invariance, were employed for the sixth and seventh subproblems, respectively (Wu et al., 2007). Difference values (Δ) between the established models were examined in interpreting CFA and multi-group CFA results in this study. ΔCFI and ΔRMSEA values were used to determine which model was more suitable for the data among the CFA results (Cheung & Rensvold, 2002), and

$\Delta\chi^2$, ΔCFI, ΔTLI, and ΔRMSEA values were considered in the invariance test. If ΔCFI, ΔTLI, and ΔRMSEA values were less than -.01 or greater than .01, the finding was interpreted as the model did not provide the relevant type of measurement invariance (Kline, 2016). In addition, the significance of the $\Delta\chi^2$ value was interpreted. Analysis assumptions before EFA and CFA were tested. EFA was performed using Factor 11.05.01 (Lorenzo-Seva & Ferrando, 2021); CFA and multi-group CFA were performed using R software (R Core Team, 2013). The psychometric properties of the forms were revealed by comparing the results based on the forms.

Results

Findings Regarding EFA and Reliability

The item factor loadings calculated from the EFA of the forms and the Cronbach's Alpha reliability coefficients for the forms are summarized below in Table 3. The Kaiser–Meyer– Olkin (KMO) values of all four forms consisting of the same items but with different category labels are higher than .90. Accordingly, the sample size is very good (Kaiser & Rice, 1974). The Bartlett Sphericity Test results of the forms were significant at .01 error level for all forms. These findings indicate the factorability of the data. In addition, the multivariate normality assumption of the forms was tested with Mardia's (1970) skewness and kurtosis coefficients, and the ULS (unweighted least squares) method was employed in the analysis. Parallel analysis was used to determine the dimensionality of the forms, and the convergence criteria for unidimensionality (UniCo, ECV, MIREAL) proposed by Ferrando & Lorenzo-Seva (2018) were considered.

Among the criteria in Table 3, the unidimensionality fit (UniCo) value is greater than .95, the explained common variance (ECV) is higher than .85, and the average of the item residual absolute loads (MIREAL) value is less than .30 indicates that the data can be handled in a unidimensional way. In light of these findings, the data sets obtained from the forms show unidimensionality. Forms 1, 2, 3, and 4 explain 76.87%, 78.65%, 79.67%, and 76.75% of the total variance with this one-dimension, respectively. The explained variance rates of the forms are very close to each other. While Form 3, with the label "Neither agree nor disagree" as the midpoint of the scale, had the highest explained variance rate, followed by Form 2 with the "I am undecided" label, Form 1 with the expression "I have no idea" at the midpoint, and Form 4, which is an end-achored scale. The similarity of the explained variance rates can be interpreted as the forms provide similar information.

Table 3
EFA and Reliability Results of the Forms

		Form 1	Form 2	Form 3	Form 4
KMO		.95	.94	.95	.95
Bartlett Sphericity Test		3805.0*	3805.0*	3805.0*	3805.0*
Explained Variance Rate		%76.87	%78.65	%79.67	%76.75
UniCo		.99	.99	.99	.99
ECV		.95	.94	.95	.93
MIREAL		.18	.20	.20	.22
Item Factor Loads	*M1*	.91	.92	.93	.91
	M2[a]	.80	.82	.83	.84
	M3	.94	.95	.96	.96
	M4	.91	.92	.92	.91
	M5	.87	.82	.87	.84
	M6	.87	.91	.89	.87
	M7	.91	.95	.94	.88
	M8	.70	.68	.71	.68
	M9	.85	.90	.90	.90
	M10	.89	.91	.89	.89
	M11	.81	.84	.85	.82
	M12	.90	.88	.90	.86
	M13[a]	.82	.82	.84	.84
	M14	.94	.93	.94	.89
Median of the Item Factor Loads		.88	.90	.89	.87
Cronbach Alpha		.97	.98	.98	.97

[a]Items added to the scale
*p< .01

According to Table 3, the item factor loads differ among the forms. Item factor loads were between .70-.94 for Form 1, .68-.95 for Form 2, .71-.96 for Form 3, and .68-.96 for Form 4. The Cronbach's alpha coefficient was calculated as .98 for Forms 2 and Form 3, and as .97 for Forms 1 and 4. Based on this finding, the reliability obtained from the forms is quite high and close to each other. While this finding is similar to the findings of Finn (1972), Wyatt & Meyers (1987), and Weng (2004), it contradicts the findings of Jacko & Huck (1974), who conducted a similar study on a multidimensional scale, and Krosnick & Berent (1993), who used seven-point scoring in their research.

Findings Regarding the Comparison of Item Factor Loads

The median values of the item factor loads in Table 3 are .88 for Form 1, .90 for Form 2, .89 for Form 3, and .87 for Form 4. Accordingly, the item factor loads obtained from all forms are quite high. The Friedman test first tested whether or not the item factor loads in Table 3 showed a significant difference from one form to

another. According to the analysis results, the item factor loads among Forms 1, 2, 3, and 4 showed a significant difference at 0.05 error level *(X^2 = 11.826; sd = 3; p = .008).*

Paired comparisons of item factor loads of the forms were evaluated using the Wilcoxon Signed Rank Test. The results of the analysis showed that the median of item factor loads differed significantly between Form 1 and 3 and between Form 3 and 4 ($p<.05$). The item factor loads did not differ significantly among other form combinations ($p>.05$). According to Cohen's (1988) criteria, the item factor loads differed between Forms 1 and 3 *(z = -2.825; p = .005; r = -.38),* and between Forms 3 and 4 with a medium effect size *(z = -2.661; p = .008; r = -.36).*

Findings Regarding the Relationship between Item Factor Loads

To answer the third sub-problem, the Spearman Correlation Coefficient was calculated to reveal the relationship among the item factor loads calculated from the forms, and the findings are presented in Table 4. According to Table 4, there is a high correlation among the item factor loads calculated from the forms. The highest correlation between item factor loads was calculated between Form 1 (I have no idea at the mid-point) and Form 3 (Neither agree nor disagree at the mid-point), and Form 2 (I am undecided at the mid-point) & Form 3 with a correlation coefficient of .95. The second highest correlation coefficient was calculated as .90 between Forms 1 and 2. Based on this finding, the item factor loads obtained from the verbal scales are highly correlated.

Table 4
Relationships Between Item Factor Loads

Forms	Form 1	Form 2	Form 3	Form 4
Form 1	1.00			
Form 2	.90*	1.00		
Form 3	.95*	.95*	1.00	
Form 4	.80*	.83*	.85*	1.00

*$p<.01$

Table 4 shows that there are high correlations in the range of .80-.85 between the item factor loads obtained from Form 4, which is in the form of an end-anchored scale, and those obtained from the verbal forms. Form 4 shows the highest correlation with factor loads obtained from Form 3, followed by Forms 2 and 1. When all the correlation coefficients in Table 4 are examined, notably lower correlation coefficients were calculated between Form 4 (the end anchored scale) and the verbal scales than between the verbal scales themselves. In other words, while the item factor loads obtained from the verbal scales had a higher correlation with each other, they showed a lower correlation with the item factor loads calculated from the end-anchored scale.

Findings Regarding Total Scores Obtained from the Forms

In the fourth sub-problem, there was no statistically significant difference among the form averages when the difference among the total scores obtained from the forms was tested with one-way ANOVA ($F_{3;1344} = 0.057$, $p>.05$). Based on this finding, the averages of the students' responses to the verbal forms in which the mid-point label differs and to the end-anchored form are similar. In other words, the student mathematical attitude scores, calculated from four different forms that were prepared for the same purpose with the same scale items but different category labels, did not show a significant difference based on the forms, and their total scores were similar. While this finding contributes to the findings of Dixon et al. (1984), Finn (1972), Wyatt and Meyers (1987), and Newstead and Arnold (1989), it contradicts those of Jacko and Huck (1974), who used a multidimensional scale in their research. The scale used in this study had a unidimensional structure. The difference between Jacko and Huck's (1974) study and this study is that a multidimensional scale was examined in their study. Dimensionality could be one of the possible reasons for this discrepancy between the two studies.

Findings Regarding Factor Scores Obtained from the Forms

In the fifth sub-problem, there was no statistically significant difference among the factor scores obtained from the forms according to one-way ANOVA ($F_{3;1344} = 0.000$; $p>.05$). On the basis of this finding, similar to the previous finding, the averages of the factor scores calculated from the verbal forms in which the mid-point label differs and from the end-anchored form are similar. In other words, the student factor scores did not show a significant difference between the forms.

Findings Regarding the Model-Data Fit

In the sixth sub-problem, the CFA results and the difference (Δ) values are presented in Table 5, below. Considering the χ^2/df and RMSEA values in Table 5, although these values indicate a low level of fit in all forms, the TLI and CFI values of the fit indices were calculated to be over .95 in all forms, which indicates that the model– data fit is at a very good level. In addition, the residuals support this goodness of model fit since SRMR has a value less than .08. Based on these findings, the CFA results provided a good level of model– data fit in all forms.

Table 5 also presents ΔRMSEA and ΔCFI over the binary combinations of all forms so that form comparisons can be made. It is noteworthy that only one of the ΔRMSEA values in Table 5 is in the range ±.01, and the other difference values exceed this limit. Chen (2007) stated that the difference is significant when ΔRMSEA is greater than .01. Therefore, it can be interpreted that there is no significant difference between Forms 2 and 3 in terms of model-data fit, but the differences between the other forms are significant. Considering Table 5, Form 1 fitted the data

better than the other forms. The item factor loads obtained as a result of CFA are given in the Appendix.

Table 5
CFA Results of the Forms and the Difference Values

Forms	χ^2	χ^2/df	RMSEA	TLI	CFI	SRMR	ΔCFI	ΔRMSEA
Form 1	306.66	3.98	.09	.99	.99	.04		
Form 2	416.51	5.41	.12	.99	.99	.05		
Form 3	437.63	5.68	.12	.98	.99	.05		
Form 4	544.35	7.07	.13	.99	.99	.06		
The difference values between Form 1 & Form 2							.00	-.02
The difference values between Form 1 & Form 3							.00	-.02
The difference values between Form 1 & Form 4							.00	-.04
The difference values between Form 2 & Form 3							.00	.00
The difference values between Form 2 & Form 4							.00	-.02
The difference values between Form 3 & Form 4							.00	-.02

Findings Regarding Measurement Invariance

Multi-group CFA was employed for the seventh subproblem. The findings regarding the extent to which Forms 1, 2, 3, and 4 provide measurement invariance are presented in Table 6. In the literature, if the difference between the Chi-square values ($\Delta\chi^2$) is insignificant and the CFI, TLI, and RMSEA difference values are within the range of ± .01, it is accepted that the relevant type of invariance is provided (Kline, 2016).

Table 6
Measurement Invariance Results

Invariance	χ^2	df	$\Delta\chi^2$	p	CFI	RMSEA	TLI	ΔCFI	ΔRMSEA	ΔTLI
Structural	328.45	308	-	-	.92	.10	.90	-	-	-
Metric	422.30	347	37.57	.54	.98	.05	.98	.06	-.05	.08
Scalar	440.48	386	47.74	.16	.98	.05	.98	-.00	-.00	.00
Strict	487.05	428	60.59	.03	.98	.04	.98	.00	-.00	.00

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation

The significance levels of $\Delta\chi^2$ values in Table 6 indicate that metric and scalar invariance can be achieved (p>.05) while strict invariance cannot (p<.05). Considering ΔCFI and ΔRMSEA values, metric invariance cannot be achieved

because these values exceed ±.01 range. When providing metric invariance, scalar invariance can be tested (Gregorich, 2006). At this point, the results obtained from scalar invariance were not interpreted , since metric invariance could not be achieved. Based on this finding, the forms do not provide metric invariance and only have structural invariance.

Discussion, Conclusion and Suggestions

The current study aimed to compare the validity and reliability evidences of the forms through the application of four different Likert-type scales consisting of the same items prepared for the same purpose. In line with this aim, an end-anchored form (Form 4) and three verbal forms (Form 1, Form 2, and Form 3) with different midpoint value labels were discussed. For the verbal forms' mid-point labels, the expressions "I have no idea" in Form 1, "I am undecided" in Form 2, and "Neither agree nor disagree" in Form 3 were included. Form 4 comprised an end-anchored form, whose beginning and end values were shown with a verbal label, and the remainder consisted of numerical values. The focus of this study was to reveal the relationship between different midpoint labels and scale versions in Likert-type scales and to present empirical evidence on the subject.

There was no significant difference between the total score averages and the factor score averages obtained from the verbal and end-anchored forms employed within the scope of the research. This finding overlaps with the findings of Dixon et al. (1984), Finn (1972), Wyatt and Meyers (1987), and Newstead and Arnold (1989). Unlike this study, Dixon et al. (1984) used a six-category (forced choice scoring) and revealed no difference between the verbal and end-anchored forms in terms of total score averages. Similarly, Finn (1972), Wyatt and Meyers (1987), and Newstead and Arnold (1989) reported that label definitions do not reveal a significant difference in the total scores in their studies, in which they compared different label formats of Likert-type scales. The current study compared five-point Likert scoring and obtained similar results in terms of total score averages. Contrary to the current findings, Jacko and Huck (1974) stated that end-anchored scales have a lower mean than verbal scales, based on the Alpert– Haber Achievement Anxiety Test (AHAAT), which is a multidimensional scale. The MAS, used as the data collection tool in the current study, has a unidimensional structure, as mentioned in the method section of this research. As Jacko and Huck (1974) stated, the non-unidimensionality of the AHAAT scale can be considered a reason for the inconsistency of the results in this context. According to the results of this study, four different forms consisting of the same items to measure the same feature presented similar information about individuals. As Newstead and Arnold (1989) noted, a significant difference between the forms in terms of total score averages would mean an increase or decrease in scale scores when using different category labels. This indicates that the total scores from the different labeled forms cannot be compared. From this perspective, it is appropriate to say that these four forms are comparable. For future studies, repeating similar studies on multidimensional and forced choice scales may be advised for researchers.

The second result obtained from the study is that the explained variance ratios and reliability of all forms are found to be close to each other as a result of EFA based on the polychoric correlation matrix. Explained variance received the highest value when the midpoint of the scale was labeled as "Neither agree nor disagree" (Form 3), followed by Form 2 labeled as "I am undecided", Form 1 labeled as "I have no idea", and end-anchored Form 4, respectively. The fact that Forms 1 and 4 have very close explained variance rates can be interpreted as these two forms provide similar information. The Cronbach's alpha internal consistency coefficients of all forms were calculated as 0.97 and 0.98, respectively. Thus, it can be concluded that the forms have similar and high reliability. While this finding is similar to the findings of the studies of Wyatt and Meyers (1987) and Weng (2004), it does not correspond to the findings of studies of Jacko and Huck (1974) and Krosnick and Berent (1993). As specified above, a multidimensional scale was used in the study of Jacko and Huck (1974). Unlike this study, Krosnick and Berent (1993) used a seven-point Likert scale to determine participants' political views. They implemented this scale in different sessions through phone calls, face-to-face interviews, and self-recording of the respondent's own responses and considered the compliance between the implementations as reliability. In this respect, the results of these two studies may not have corresponded. Within this scope, future studies can examine the compliance between different implementations (for example, face-to-face and online) and compare the results of EFA and CFA.

This study showed that there was a significant difference between factor loadings obtained from EFA results with medium effect size when the mid-point value was labeled as "I have no idea" & "Neither agree nor disagree.", and "Neither agree nor disagree." and end-anchored one. Moreover, regarding the correlations between the item factor loads obtained from the forms, it was concluded that the item factor loads obtained from the verbal forms had a higher correlation with each other; however, they showed a lower correlation with the item factor loads obtained from the end-anchored form. In other words, the forms consisting of verbal attitude labels have a higher level of correlation between the item factor loads and lower relationships with the end-anchored scale. Thus, the numerical– verbal statement difference in attitude labels of scales changes individuals' responses relating to the characteristic to be measured. Based on this result, individuals perceive different attitude labels in the same category in different ways. Researchers may be advised to conduct similar studies using scales measuring different characteristics.

As an interesting result of the study, it is notable that the limit is not exceeded in only one comparison according to the ΔRMSEA values. There is no significant difference in terms of model-data fit between Form 2, with the expression at the mid-point "I am undecided" and Form 3, which has the label "Neither agree nor disagree". There were significant differences with respect to the ΔRMSEA in the other form combinations in which the model– data fit was compared, and among the forms, Form 1 with the label "I have no idea" fitted the data better. Based on this, placing the "I have no idea" label at the midpoint is more appropriate for the students' responses

compared with the other verbal forms and end-anchored scale. However, according to Başar (2001), statements such as “neutral, no idea, no comment” do not convey a medium degree attitude level, but a status. Nevertheless, according to Nadler et al. (2015), a medium degree in attitude statements means "neutral" and includes the tendency toward the center and the desire for social approval of the individuals filling out the form. Therefore, the model– data fit of Form 1 may have been better.

According to the results obtained in terms of measurement invariance, only structural invariance was achieved between the forms measuring the same characteristic and having the same scale items with different category labels, but metric invariance was not achieved. The lack of metric invariance across the forms indicates that item factor loads differ across the forms. In this case, it can be examined whether there is item bias (DIF) or item parameter shift (DRIFT) in the items according to the forms using quantitative and/or qualitative methods. Considering the findings of this study, researchers can repeat such studies as they are limited in number, examine the validity and reliability of verbal , end-anchored, and numerical forms, compare the factor loads of the scale items with different techniques, and determine the sources of variability based on the generalizability theory and item response theory.

The data collection step of this study was conducted on an online platform due to the COVID-19 pandemic. During the data collection phase, the links of Forms 1, 2, 3, and 4 were kept accessible for students for four days at one-week interval. To further explain this process, Form 1 was first administered to the students. The link to this form was accessible to students for four days and was blocked after four days. Exactly one week after access to Form 1 was blocked, the link to Form 2 was distributed to the students and the same timing was respected for all forms. The order of form application was kept consistent in this study, considering that data collection would be difficult in the pandemic and may be confusing for students, considering the different priority and posteriority combinations of the forms. In this case, errors arising from the sequence effect may affect the measurement process. This appears to be a limitation of this study. For future studies, it may be suggested to include different priority– posteriority combinations of forms to eliminate the sequence effect in the results. Moreover, the scale used in this study is the MAS, which is scored on a five-point Likert-type scale, consists of 14 items, and has a unidimensional structure. It may be recommended to repeat this type of study on scales that contain more or fewer items, are scored in different types such as three-point or four-point Likert, and show multidimensionality. In addition, in this study, only the mid-point labels were changed in the verbal scales, whereas the other four labels remained the same. The effect of different label statements on attitude can also be examined by changing all label statements or sorting the labels in different ways.

Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi
Yıl: 2024, Cilt: 57, Sayı:1, 251-288
DOI: 10.30964/auebfd.1280498, E-ISSN: 2458-8342, P-ISSN: 1301-3718

Farklı Şekilde Etiketlenen Likert Tipi Ölçeklerin Ölçme Değişmezliğine Etkisinin İncelenmesi[1]

MAKALE TÜRÜ	Başvuru Tarihi	Kabul Tarihi	Yayım Tarihi
Araştırma Makalesi	16.04.2023	16.07.2023	06.03.2024

Nuri Doğan [2]
Hacettepe Üniversitesi

Ceylan Gündeğer Kılcı [3]
Aksaray Üniversitesi

Meltem Yurtçu [4]
İnönü Üniversitesi

Öz

Bu araştırma, farklı orta nokta etiketine sahip üç sözel ve bir ucu gömülü ölçek formunun psikometrik özelliklerini incelemeyi amaçlamaktadır. Bu amaçla, 377 üniversite öğrencisinden, aynı özelliği ölçen ve aynı maddelerden oluşan fakat farklı tutum etiketlerine sahip dört farklı ölçek formu ile veri toplanmıştır. Araştırmada, tek boyutlu bir yapıya sahip olan Matematiksel Tutum Ölçeği'nin 14 maddelik kısa formu kullanılmıştır. Veri seti, geçerlik, güvenirlik ve ölçme değişmezliği bakımından incelenmiştir. Sonuçlar, dört formun açıklanan varyans oranlarının yakın olduğunu, formların benzer ve yüksek düzeyde güvenirliğe sahip olduğunu göstermiştir. Doğrulayıcı Faktör Analizi sonuçlarına göre, orta nokta etiketi "Fikrim yok" olan Form 1, veriye daha iyi uyum sağlamıştır. Son olarak, formlar yalnızca yapısal değişmezliği sağlamıştır. Metrik değişmezliğin sağlanamaması madde faktör yüklerinin formlar arasında değiştiğine işaret etmektedir. Bu sonuca dayanarak bireylerin aynı kategoride yer alan farklı tutum etiketlerini ve farklı türdeki ölçekleri farklı şekillerde algıladıkları söylenebilir. Bu bulgular ışığında, araştırmacılar genellenebilirlik kuramı ve/veya madde tepki kuramı temelinde benzer çalışmaları tekrarlayarak alana deneysel kanıtlar koyabilirler.

Anahtar sözcükler: Likert, etiket, kararsızım, fikrim yok, ne katılıyorum ne katılmıyorum, ucu gömülü, ölçek.

[1]Bu araştırma, 1-4 Eylül 2021 tarihleri arasında Türkiye'de çevrimiçi düzenlenen 7. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde (CMEEP-2020) sözlü bildiri olarak sunulmuştur.

[2]Prof. Dr., Hacettepe Üniversitesi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, E-posta: nuridogan2004@gmail.com, https://orcid.org/0000-0001-6274-2016

[3]*Sorumlu Yazar*: Dr. Öğr. Üyesi, Aksaray Üniversitesi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, E-posta: cgundeger@gmail.com, https://orcid.org/0000-0003-3572-1708

[4]Doç. Dr., İnönü Üniversitesi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, E-posta: meltem.yurtcu@gmail.com, https://orcid.org/0000-0003-3303-5093

Kişilik ve tutum gibi psikolojik özelliklerin ölçülmesinde sık kullanılan tekniklerden biri dereceleme toplamları tekniğidir. Bu yaklaşım, yaygın olarak Likert yöntemi olarak adlandırılmaktadır (Anastasi, 1982). Likert tipi ölçeklerde, cevaplayıcı, ölçekteki her ifadenin kapsadığı tutum ögesine katılma veya katılmama derecesini bildirmektedir (Tezbaşaran, 2008). Seashore ve Katz (1982), Likert tekniğinin görünüşte kolay ancak uygulamada oldukça karmaşık olduğunu ifade etmiştir. Teknik, tutum ifadelerinin hazırlanması ve yanıtlanması bakımından kolaylık sağlamaktadır. Tekniğin karmaşıklığı ise tek boyutlu ölçekler elde etme, tutumların boyutluluğunu keşfetme, ölçülen değişkenler arası yapısal ve nedensel ilişkileri değerlendirme amacıyla madde korelasyonu, faktöryel ve çok boyutlu analiz prosedürlerini kullanmada yatmaktadır (Seashore ve Katz, 1982). Likert tipi ölçeklerde sıklıkla kullanılan tutum etiketlerinin formatları, *sözel* (verbal) ölçekler, *ucu gömülü* (end anchored) ölçekler ve *sayısal* (numerical) ölçekler olmak üzere üçe ayrılmaktadır. Sözel ölçeklerde, tüm ölçek noktalarının sözel bir ifadeyle etiketlendirilmesi söz konusudur. Ucu gömülü ölçeklerin ilk ve son ölçek noktaları sözel etikete sahip ve geri kalan kısım sayısaldır. Sayısal ölçekte ise tüm ölçek noktaları sayısal değere sahiptir ve bazen bu ölçek noktaları belirli yüzdelik etiketleri (%0, %25 vb.) ile adlandırılabilmektedir (Newstead ve Arnold, 1989).

Likert tipi ölçeklerde genellikle *Kesinlikle Katılıyorum*'dan (5) *Kesinlikle Katılmıyorum*'a (1) kadar beş kategori yer almaktadır (Jamieson, 2004). Beşli Likert tipi sözel ölçek etiketleri, *Kesinlikle katılmıyorum, Katılmıyorum, Kararsızım, Katılıyorum ve Kesinlikle katılıyorum* şeklinde sıralansa da alanyazında bu sıralamanın ve etiketlerin nasıl olacağına ilişkin fazla bilgi yer almamaktadır. Örneğin Robie ve diğ. (2022) tarafından yürütülen yeni bir araştırmada kategorilerin farklı sıralandığı durumlarda ölçme değişmezliğinin, ölçek ortalamasının ve cevaplayıcı tepkisinin pek değişmediği görülmüştür. Benzer şekilde, cevaplayıcının bilgisi olmaması durumunu temsil etmesi açısından "Bilmiyorum" seçeneğinin sunulmasının önerildiği bazı araştırmalara rastlansa da (Payne, 1950; Vaillancourt, 1973) bu seçeneğin güvenirliği güçlendirdiğini söylemek pek mümkün değildir (Krosnick ve Presser, 2010). Bu sebeple araştırmacılar ölçek geliştirme aşamasında ölçek kategorilerinin sıralanışı ve etiketler hakkındaki bu kararı, kendi bilgi ve deneyimlerine uygun şekilde kendileri vermektedir. Seçeneklerin sıralanışı ve etiketlerin yanında özellikle *ölçek orta noktası* (midpoint) olarak çalışmalarda çoğunlukla, İngilizcedeki "Undecided" ifadesinin Türkçe karşılığı "Kararsızım" ifadesinin kullanıldığı belirtilmektedir (Kağıtçıbaşı, 2010). Bazı araştırmalarda orta seçenek, "Ne katılıyorum ne katılmıyorum" (Gegez, 2010; Kurtuluş, 2006; Yükselen, 2003); bazı çalışmalarda ise "Fikrim yok" ifadesi ile etiketlenmiştir (Nakip, 2006). Ölçeğin güvenirliğinin yüksek olabilmesi için ölçek noktalarındaki etiketlerin açık ve anlaşılır olması bir gereklilik olarak karşımıza çıkmaktadır. Eğer etiketlerde belirsiz ifadeler yer alıyorsa, ölçeğin geçerliği ve güvenirliği tehlikeye düşebilmektedir (Krosnick ve Presser, 2010).

Alanyazında, tutum etiketlerinin gösterdikleri derecelerin birbirine benzer olup olmadığı; dil açısından (semantik) aralarında farklılık olup olmadığı; bireyler

tarafından bu ifadelerin eş-anlamlı olarak anlaşılıp anlaşılmayacağı hakkında bir görüş birliği bulunmamaktadır. Geçmişten günümüze, özellikle orta değerin ne bildirdiğine ilişkin çeşitli tartışmalar gündeme gelmiştir. Başar'a (2021) göre, "Kararsızım, Fikrim yok, Bir şey söyleyemem" gibi ifadeler orta derece veya eğilim değil, durum bildirmektedir; dolayısıyla bu tür ifadelerin ölçek uygulamalarında kullanılmaları hatalı olabilmektedir. Başar (2021) bu ifadelerin yerine *"*Yansızım*", "*Nötrüm*", "*Orta düzeyde katılıyorum*"* gibi ifadeye katılma ve katılmama bildirebilecek bir etiket kullanılması gerektiğini belirtmektedir. Bununla birlikte alanyazın incelendiğinde Likert (1967) ve Bogardus'un (1967) daha sonraki çalışmalarında orta seçenekte "Undecided" sözcüğüne yer vermediği; Thurstone (1967) ile Allen ve Kenney'nin (1967) orta seçenek için "Neutral" (yansız) sözcüğünü kullandığı görülmektedir (akt., Başar, 2021, s.3). Tezbaşaran (2008) ile Turgut ve Baykul (1992) da "Kararsızım" sözcüğünü "ortadayım, yansızım, nötrüm, fikrim yok" anlamında kullandıklarını belirtmiştir. Başar'a göre (2021) nötr, yansız anlamı taşır ve bir karar bildirir. *Kararımı verdim: Herhangi bir yanda değilim, ortadayım.* "Kararsızım" ise karar bildirmez, kararın ne diye soran araştırmacıya, kararım yok, der. Kararı olmayana, karar derecesi bildiren puan verilemez (Başar, 2021). Bora Semiz ve Altunışık (2016), orta değer için *"*Kararsızım*", "*Ne katılıyorum ne katılmıyorum*"* etiketlerinin kullanılabileceğini ancak "Fikrim yok" ifadesinin ölçek orta noktası için tutumda bir yer belirtmediğinden orta nokta için uygun olmayacağını belirtmiştir. Ancak bir başka çalışma göstermiştir ki *"*Nötr*"* anlamında kullanılan orta nokta seçeneğine cevaplayıcılar, merkeze yönelme yanlılığı ve sosyal beğenilme arzusu nedeniyle de yönelebilmektedir (Nadler ve diğ., 2015).

Ölçek etiketlerindeki farklılıklar, ölçeklerden elde edilen toplam puan, ölçeklerin güvenirlikleri ve geçerlikleri ile ölçek cevaplarındaki değişkenlik üzerinde çeşitli farklılıklar ortaya çıkarabilmektedir (Newstead ve Arnold, 1989). Alanyazın incelendiğinde, Dixon ve diğ. (1984) tarafından yürütülen bir araştırmada, sözel ve ucu gömülü Likert tipi ölçeklerden elde edilen puanların etiketlere göre anlamlı bir farklılık göstermediği sonucuna ulaşılmıştır. Benzer şekilde Finn (1972) ile Wyatt ve Meyers (1987) de, Likert tipi ölçeklerin farklı etiket formatlarını karşılaştırdığı çalışmalarında, etiket tanımlarının puanlar üzerinde anlamlı bir farklılık ortaya çıkarmadığını; farklı etikete sahip ölçeklerin güvenirliklerinin ise benzer olduğunu göstermiştir. Jacko ve Huck (1974) ise bu araştırmaların tersine, ucu gömülü ölçeklerin sözel ve sayısal ölçeklere kıyasla daha düşük ortalamaya sahip olduğunu ortaya koymuştur. Benzer şekilde, Krosnick ve Berent (1993) ile Weng (2004) de ucu gömülü ölçeklerin güvenirliğinin sözel ölçeklerin güvenirliğine kıyasla daha düşük olduğunu ifade etmişlerdir. Sayısal ölçekler üzerinden yürütülen araştırmalara göz atıldığında ise Blumberg ve diğ. (1966) ucu gömülü ölçekler ve sayısal ölçekler arasında bir farklılık olmadığını ortaya koymuştur. Peter ve McCormick (1966) sözel ölçeklerin sayısal ölçeklerden daha yüksek güvenirliğe sahip olduğunu göstermiştir (akt., Newstead ve Arnold, 1989, s.35). Newstead ve Arnold (1989) ise çalışmalarında sayısal ölçeklerin ucu gömülü ölçeklere kıyasla daha yüksek ortalamaya sahip

olduğunu ancak sözel ve ucu gömülü ölçekler arasında ortalama bakımından manidar bir farklılık olmadığını belirtmiştir.

Alanyazında Likert tipi ölçekler çoğunlukla ölçeklerdeki kategori sayısı temel alınarak güvenirlik açısından karşılaştırılmıştır. Bazı araştırmalarda kategori sayısı arttıkça güvenirliğin arttığı görülürken (Alwin, 1992; Bandalos ve Enders, 1996; Hartley ve MacLean, 2006; Kılıç, Uysal ve Kalkan, 2021; Lee ve diğ., 2002; Simms ve diğ., 2019); bazı araştırmalarda güvenirliğin kategori sayısına bağlı şekilde anlamlı bir farklılık göstermediği ancak beş kategorinin en uygun (optimal) güvenirliği verdiği ortaya koyulmuştur (Finn, 1972). Ayrıca kimi çalışmalarda ölçek orta noktası olmayan (forced-choice) çift sayılı kategori ile daha yüksek güvenirlik elde edilirken (Bendig, 1954); kimi bilim insanları ölçeğe orta nokta eklenmesiyle güvenirliğin güçleneceğini belirtmiştir (O'Muircheartaigh ve diğ., 2000). Özetle, farklı tutum etiketlerine sahip sözel ölçekler ile ucu gömülü ölçeklerin geçerlik ve güvenirliklerinin incelendiği sınırlı sayıda araştırma olduğu ve yapılan araştırmaların birbiriyle tutarlı sonuçlar vermedikleri görülmektedir. Aynı amaca yönelik, aynı ölçek maddelerinden oluşan fakat farklı etiketlere sahip ölçeklerin ölçme değişmezliğini ne düzeyde sağladığının ise alanyazında pek incelenmediği söylenebilir.

Ölçme değişmezliği, ölçülen özellik bakımından eşit düzeyde olan grupların ölçme aracından aldıkları ham puanın aynı olması ya da farklı gruplarda yer alan bireylerin ölçme aracındaki maddeleri aynı şekilde algılaması ve yorumlaması anlamına gelmektedir (Bryne ve Watkins, 2003.) Gregorich'e (2006) göre, ölçme değişmezliğinde hiyerarşik dört değişmezlik türünden söz edilmektedir. Bunlardan ilki yapısal (şekilsel) değişmezliktir. Yapısal değişmezlikte, ölçeğin faktör yapısının gruplarda eşitliği sınanmaktadır. Bu değişmezliğin sağlanması durumunda metrik (zayıf) değişmezlik test edilebilmektedir. Metrik değişmezlikte grupların ölçek maddelerini benzer şekilde algılayıp algılamadığı kontrol edilir. Metrik değişmezliğin sağlanamaması madde faktör yüklerinin gruplarda eşit olmadığı anlamına gelmektedir. Dolayısıyla metrik değişmezlik sağlanamazsa gruplardan elde edilen puanların karşılaştırılması yanlı olabilmektedir. Metrik değişmezliğin sağlanması durumunda skalar (güçlü) değişmezlik sınanmaktadır. Skalar değişmezlikte, madde faktör yükleri ile regresyon sabitlerinin gruplar arasında eşit olup olmadığı incelenmektedir. Skalar değişmezliğin sağlanması durumunda katı değişmezlik ele alınmakta ve katı değişmezlik ile de hata varyanslarının gruplarda eşitliği test edilmektedir. Sıralanan dört değişkenlik türü birbirinin ön koşulu olacak şekilde ele alınmaktadır. Bu durumda yapısal değişmezliğin sağlanamaması diğer değişmezliklerin de sağlanmadığı anlamını taşır. Dolayısıyla yapısal değişmezliğin sağlanamadığı bir ölçme aracından elde edilen puanların karşılaştırılması da yanlılık içerecektir.

Bu çalışmada, aynı öğrenci grubundan elde edilen veriyle, aynı maddelerden oluşan fakat farklı kategori etiketlerine sahip ölçek formlarının (sözel ve ucu gömülü beşli Likert tipi ölçeklerin) geçerlik ve güvenirlik kanıtlarını incelemek, bu kanıtları formlar üzerinden karşılaştırmak ve formların ölçme değişmezliğini ne düzeyde

sağladığını ortaya çıkarmak amaçlanmıştır. Alanyazındaki sınırlı sayıda çalışma ile ölçeklerdeki kategori etiketlerinin farklı kullanımlarına ilişkin ampirik bir kanıt sunması bakımından bu araştırmanın önemli ve gerekli olabileceği düşünülmüştür. Araştırmada aşağıdaki alt problemlere cevap aranmıştır:

1. Farklı kategori etiketlerine sahip ölçeklerin Açımlayıcı Faktör Analizi (AFA) sonucu hesaplanan madde faktör yükleri ve Cronbach Alfa güvenirlik katsayıları nasıldır?
2. Farklı kategori etiketlerine sahip ölçeklerin AFA sonucu hesaplanan madde faktör yükleri arasında nasıl bir farklılık vardır?
3. Farklı kategori etiketlerine sahip ölçeklerin AFA sonucu hesaplanan madde faktör yükleri arasında nasıl bir ilişki vardır?
4. Farklı kategori etiketlerine sahip ölçeklerden elde edilen toplam puanlar nasıl bir farklılık göstermektedir?
5. Farklı kategori etiketlerine sahip ölçeklerden elde edilen faktör puanları nasıl bir farklılık göstermektedir?
6. Farklı kategori etiketlerine sahip ölçeklerin Doğrulayıcı Faktör Analizi sonucu hesaplanan uyum indeksleri arasında nasıl bir farklılık vardır?
7. Farklı kategori etiketlerine sahip ölçekler ölçme değişmezliğini ne düzeyde sağlamaktadır?

Yöntem

Araştırmanın bu bölümünde araştırma modeline, çalışma grubuna, veri toplama araçlarına, etik kurul kararına, verilerin toplanma sürecine ve verilerin analizine yer verilmiştir.

Araştırma Modeli

Bu araştırma, aynı maddelerden oluşan ve farklı tutum etiketlerine sahip ölçek formlarının geçerlik ve güvenirliklerinin incelenmesi bakımından betimsel araştırma niteliğindedir. Betimsel araştırmalarda "… nedir?" sorusuna cevap aranmaktadır (Balcı, 2011). Ayrıca farklı formlardan hesaplanan madde faktör yükleri arasındaki ilişkinin ortaya çıkarılması bakımından da araştırmanın korelasyonel olduğu söylenebilir. Korelasyonel araştırmalarda korelasyon katsayısı ile ilişkinin derecesi ortaya çıkarılmaya çalışılır (Balcı, 2011).

Çalışma Grubu

Araştırmanın çalışma grubunu, 2020-2021 eğitim öğretim yılında, Türkiye'de farklı üniversitelerde lisans düzeyinde öğrenim görmekte olan 377 öğrenci oluşturmaktadır. Çalışma grubuna ilişkin demografik bilgiler Tablo 1'de yer almaktadır. Tablo 1'e göre çalışma grubunda yer alan öğrencilerin büyük bir kısmını

eğitim fakültesi öğrencileri oluşturmaktadır. Katılımcıların önemli bir kısmı ikinci ve üçüncü sınıf düzeyinde öğrenim görmekte olup çoğu kız öğrencilerden oluşmaktadır.

Tablo 1
Çalışma Grubunun Demografik Özellikleri

Özellik	Kategori	n	%
Cinsiyet	Kız	272	72.1
	Erkek	105	27.9
	Toplam	377	100
Fakülte	Eğitim Fakültesi	346	91.8
	Spor Fakültesi	25	6.6
	Diğer	6	1.6
	Toplam	377	100
Sınıf Düzeyi	Birinci Sınıf	57	15.1
	İkinci Sınıf	138	36.6
	Üçüncü Sınıf	161	42.7
	Dördüncü Sınıf	21	5.6
	Toplam	377	100

Veri Toplama Araçları

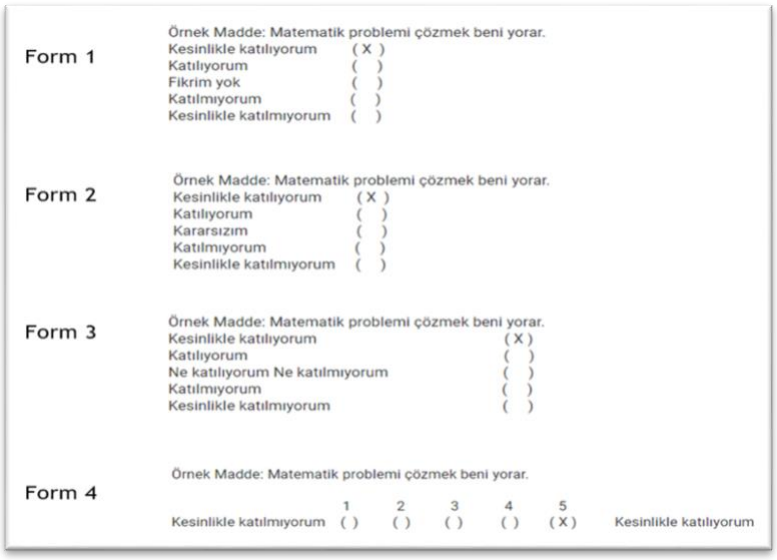
Araştırmada veri toplama aracı olarak Matematikle İlgili Düşünceler Ölçeği (MİDÖ) kullanılmıştır. MİDÖ, Baykul (1990) tarafından geliştirilmiş, 15'i olumlu ve 15'i olumsuz ifadeden olmak üzere toplam 30 maddeden oluşan tek boyutlu bir ölçektir. MİDÖ'nün tek boyutla açıklanan varyans oranı %56 ve güvenirlik katsayısı da .96'dır. Ölçekten alınabilecek en düşük puan 30 ve en yüksek puan 150'dir (akt., Nartgün, 2002, s.47).

Bu araştırma kapsamında, 30 maddelik bu ölçeğin dört ayrı formunun öğrencilere uygulanmasının zaman alacağı düşünülerek ve ölçekteki bazı maddelerin yeterli bulunamaması veya çok spesifik olması sebebiyle ölçek kısaltılmıştır. Bu çalışmada amaç öğrencilerin ölçek puanlarına ilişkin bir çıkarım yapmak olmadığından ölçeğin kısaltılmasının yapı geçerliğine olan etkisi de göz önünde bulundurularak araştırma kapsamında geçerlik ve güvenirlik kanıtları toplanmıştır. Ölçeğin kısaltılmasında, doğrudan Matematik dersine yönelik düşüncelere işaret eden ifadelerin çıkarılmasına (*örneğin, Matematik çok sevdiğim dersler arasındadır.)*, bununla birlikte Nartgün'de (2002) madde faktör yükü yüksek şekilde elde edilen maddelerin ölçekte kalmasına dikkat edilmiştir. Kısaltılan ölçeğe, matematiğe yönelik düşünceleri ölçtüğü düşünülen daha ilgi çekici iki madde eklenmiştir. Eklenen maddeler, *"Bir durumu matematiksel olarak ifade etmek beni mutlu eder."* ve *"Matematiksel keşifler beni büyüler."* şeklindedir. Öğrencilere uygulanan ölçekte yer alan madde sayısı 14'tür. Bunların altısı Matematiğe yönelik olumsuz düşünceleri; sekizi ise olumlu düşünceleri temsil etmektedir. Ölçeğin yalnızca tutum etiketleri farklılaştırılarak dört farklı versiyonu oluşturulmuştur. Form 1, Form 2, Form 3 (sözel formlar) ve Form 4 (ucu gömülü form) olarak adlandırılan bu versiyonlara ilişkin tutum etiketlerindeki farklılık Şekil 1'de sunulmuştur.

Şekil 1'de görüldüğü üzere, ilk üç sözel formda yalnızca orta düzeyi gösteren etiket değiştirilmiştir. Orta düzey için Form 1'de "Fikrim Yok"; Form 2'de "Kararsızım" ve Form 3'te "Ne katılıyorum ne katılmıyorum" ifadeleri yer almıştır. Form 4'te ise ilk ve son değerleri sözel etiketle gösterilen ve geriye kalan kısmı sayısal değerlerden oluşan ucu gömülü bir ölçek kullanılmıştır.

Şekil 1

Uygulanan Formlar



Etik Kurul Kararı

Veri toplama aşaması öncesinde, İnönü Üniversitesi Bilimsel Araştırmalar ve Etik Kurulu'ndan gerekli izinler alınmıştır (Protokol No: 13-19, Tarih: 02/07/2021). Ayrıca araştırmaya katılımda sözlü onay istenerek gönüllülük esas alınmıştır.

Verilerin Toplanma Süreci

Veri, çevrimiçi platformda ve birer hafta arayla ölçek uygulamalarını içerecek şekilde toplanmıştır. Veri toplama sürecinde ölçek formlarının linkleri öğrencilerle paylaşılmış, ölçeği doldurmaları için verilen süre (4 günün) sonunda link öğrencilere kapatılmıştır. Böylece öğrencilerin formları daha önce ya da daha sonra görmelerinin önüne geçilmiştir. Formlar arası eşleşmenin yapılabilmesi adına öğrencilerden TC kimlik numarasının ilk altı hanesini yazmaları istenmiştir. Veri toplama süreci sonunda bu bilgiler kontrol edilmiş, tekrar eden (bir formu birden fazla kez dolduran) bireylere ait veriler tespit edilerek veri setinden çıkarılmıştır.

Verilerin Analizi

Veri analizinden önce, faktör analizi varsayımlarının test edilmesi aşamasında tek yönlü ve/veya çok yönlü uç değer gösteren toplam 40 kişi analiz dışında bırakılmıştır. Formların tümünü tekrarsız yanıtlayan 337 kişiye ait betimsel istatistikler Tablo 2'de yer almaktadır. Tablo 2'deki çarpıklık ve basıklık katsayılarına göre cevapların normal dağılımdan aşırı sapma göstermediği söylenebilir. Değişkenlere ait ortalama ve ortanca değerleri birbirine yakındır. Formlardan elde edilecek en düşük ve en yüksek değerler puan rajının tüm tutum düzeylerinin kapsadığına işaret etmektedir. Standart sapma değerleri incelendiğinde ilk üç sözel formda bu değerlerin benzer olduğu, dördüncü (ucu gömülü) formda ise değerin bir miktar yükseldiği görülmektedir. Tablo 2'ye göre formların birbirine benzer bilgiler verdiği söylenebilir.

Tablo 2
Formların Betimsel İstatistikleri

İstatistikler	Form 1	Form 2	Form 3	Form 4
N	337	337	337	337
Ortalama	45.98	45.89	45.63	46.08
Ortanca	48.00	47.00	47.00	48.00
Tepedeğer	56.00	55.00	55.00	66.00
Standart Sapma	14.61	14.47	14.62	15.80
Çarpıklık Katsayısı	-0.33	-0.30	-0.31	-0.30
Standart Hata	0.13	0.13	0.13	0.13
Basıklık Katsayısı	-0.91	-0.87	-0.85	-1.02
Standart Hata	0.27	0.27	0.27	0.27
En Düşük Değer	14.00	14.00	14.00	14.00
En Yüksek Değer	70.00	70.00	70.00	70.00

Birinci alt problemin çözümünde, veri setine polikorik korelasyon matrisine dayalı Açımlayıcı Faktör Analizi (AFA) uygulanmış ve ölçeğin tek boyutluluk özelliği göstermesi sebebiyle formlar için Cronbach Alfa iç tutarlılık katsayısı hesaplanarak formların güvenirlikleri belirlenmiştir. İkinci alt problemin çözümünde, AFA sonucu hesaplanan madde faktör yükleri arasındaki farklılık Friedman ve Wilcoxon testleri ile sınanmış ve testlere ilişkin etki büyüklükleri hesaplanmıştır. Üçüncü alt problemin çözümü için madde faktör yükleri arasındaki ilişki Spearman Korelasyon Katsayısı yardımıyla incelenmiştir. Dördüncü ve beşinci alt problemde tek yönlü varyans analizi ile formlardan elde edilen toplam puanlar ve faktör puanlarının formlar arasında farklılık gösterip göstermediği incelenmiştir.

Altıncı alt problemde Doğrulayıcı Faktör Analizi (DFA) ve yedinci alt problemde ise ölçme değişmezliğinin sınanması amacıyla sıklıkla başvurulan bir

yöntem olan çoklu grup DFA işe koşulmuştur (Wu ve diğ., 2007). Bu araştırmada DFA ve çoklu grup DFA analiz sonuçlarını yorumlamada, kurulan modeller arasındaki fark (Δ) değerleri incelenmiştir. DFA sonuçları arasında hangi modelin veriye daha uygun olduğunu belirlemede ΔCFI ve ΔRMSEA değerleri (Cheung ve Rensvold, 2002); değişmezlik testinde ise $\Delta\chi^2$, ΔCFI, ΔTLI ve ΔRMSEA değerleri dikkate alınmıştır. ΔCFI, ΔTLI ve ΔRMSEA değerleri, -.01'den küçük veya .01'den büyük ise, bulgu, modelin ilgili değişmezlik türünü sağlamadığı şeklinde yorumlanmıştır (Kline, 2016). $\Delta\chi^2$ değerinin ise manidarlığı yorumlanmıştır. AFA ve DFA öncesi analiz varsayımları test edilmiştir. AFA, Factor 11.05.01 programında (Lorenzo-Seva ve Ferrando, 2021); DFA ve çoklu grup DFA, R yazılımında (R Core Team, 2013) gerçekleştirilmiştir. Elde edilen sonuçlar formlar temelinde karşılaştırılarak formların psikometrik özellikleri ortaya çıkarılmaya çalışılmıştır.

Bulgular

AFA ve Güvenirliğe İlişkin Bulgular

Formların Açımlayıcı Faktör Analizi (AFA) sonucu hesaplanan madde faktör yükleri ve formlara ait Cronbach Alfa güvenirlik katsayıları aşağıda Tablo 3'te özetlenmiştir. Aynı maddelerden oluşan fakat farklı kategori etiketlerine sahip formların dördünün de Kaiser-Meyer-Olkin (KMO) değerleri .90 değerinden yüksektir. Buna göre formlar için örneklem büyüklüğünün çok iyi düzeyde olduğu söylenebilir (Kaiser ve Rice, 1974). Formların Bartlett Küresellik Testi sonuçları ise tüm formlarda .01 hata düzeyinde manidar çıkmıştır. Bu bulgular verinin faktörleşebilirliğine işaret etmektedir. Bununla birlikte formların çok değişkenli normallik varsayımı Mardia'nın (1970) çarpıklık ve basıklık katsayıları ile test edilmiş ve analizde ULS (ağrılıklandırılmamış en küçük kareler) yöntemi işe koşulmuştur. Formların tek boyutluluğuna karar vermede paralel analizden yararlanılmış ve Ferrando & Lorenzo-Seva (2018) tarafından önerilen tek boyutluluğa yakınsama kriterleri (UniCo, ECV, MIREAL) dikkate alınmıştır.

Tablo 3'te yer alan kriterlerden tek boyutluluğa uyum (UniCo) değerinin .95'ten büyük olması, açıklanan ortak varyans (ECV) değerinin .85'ten yüksek olması ve madde artık mutlak yüklerinin ortalaması (MIREAL) değerinin .30'dan küçük olması, verinin tek boyutlu şekilde ele alınabileceğine işaret etmektedir. Bu bulgular ışığında formlardan elde edilen veri setlerinin tek boyutluluk özelliği gösterdiği söylenebilir. Form 1, Form 2, Form 3 ve Form 4, bu tek boyutla toplam varyansın sırasıyla %76.87; %78.65; %79.67 ve %76.75'ini açıklamaktadır. Formlara ait açıklanan varyans oranları birbirine oldukça yakındır. Ölçek orta değeri olarak "Ne katılıyorum ne katılmıyorum" etiketinin yer aldığı Form 3 en yüksek açıklanan varyans oranına sahipken; bunu takiben sırasıyla orta değerde *"*Kararsızım*"* etiketinin yer aldığı Form 2; orta değerde *"*Fikrim yok*"* ifadesinin yer aldığı Form 1 ve son olarak ucu gömülü ölçek olan Form 4 gelmektedir. Açıklanan varyans oranlarının yakınlığı, formların benzer düzeyde bilgi verdiği şeklinde yorumlanabilir.

Tablo 3
Formlara Ait AFA ve Güvenirlik Bulguları

		Form 1	Form 2	Form 3	Form 4
KMO		.95	.94	.95	.95
Bartlett Küresellik Testi		3805.0*	3805.0*	3805.0*	3805.0*
Açıklanan Varyans Oranı		%76.87	%78.65	%79.67	%76.75
UniCo		.99	.99	.99	.99
ECV		.95	.94	.95	.93
MIREAL		.18	.20	.20	.22
Madde Faktör Yükleri	*M1*	.91	.92	.93	.91
	M2[a]	.80	.82	.83	.84
	M3	.94	.95	.96	.96
	M4	.91	.92	.92	.91
	M5	.87	.82	.87	.84
	M6	.87	.91	.89	.87
	M7	.91	.95	.94	.88
	M8	.70	.68	.71	.68
	M9	.85	.90	.90	.90
	M10	.89	.91	.89	.89
	M11	.81	.84	.85	.82
	M12	.90	.88	.90	.86
	M13[a]	.82	.82	.84	.84
	M14	.94	.93	.94	.89
Madde Faktör Yükleri Ortancası		.88	.90	.89	.87
Cronbach Alfa		.97	.98	.98	.97

[a]Ölçeğe eklenen maddeler
*p<.01

Tablo 3 incelendiğinde, madde faktör yüklerinin formlar arasında farklılaştığı görülmektedir. Madde faktör yükleri Form 1 için .70-.94; Form 2 için .68-.95; Form 3 için .71-.96; Form 4 için .68-.96 ranjındadır. Formların Cronbach Alfa katsayıları ise Form 2 ve Form 3 için .98; Form 1 ve Form 4 için ise .97'dir. Buna dayanarak formlardan elde edilen güvenirliklerin oldukça yüksek ve birbirine yakın olduğu söylenebilir. Bu bulgu, Finn (1972), Wyatt ve Meyers (1987) ile Weng (2004) ile benzerlik gösterirken; çok boyutlu bir ölçek üzerinden benzer bir araştırma yürüten Jacko ve Huck (1974) ile araştırmalarında yedili puanlama kullanan Krosnick ve Berent'in (1993) çalışmaları ile örtüşmemektedir.

Madde Faktör Yüklerinin Karşılaştırılmasına İlişkin Bulgular

Tablo 3'te yer alan madde faktör yüklerinin ortanca değerleri Form 1 için 0.88; Form 2 için 0.90; Form 3 için 0.89 ve Form 4 için 0.87 şeklinde hesaplanmıştır. Buna göre tüm formlardan elde edilen madde faktör yüklerinin oldukça yüksek olduğu söylenebilir. Tablo 3'te yer alan madde faktör yüklerinin formdan forma manidar bir değişim gösterip göstermediği öncelikle Friedman testi ile sınanmıştır. Analiz

sonuçlarına göre Form 1, Form 2, Form 3 ve Form 4 arasında madde faktör yüklerinin .05 hata düzeyinde anlamlı bir farklılık gösterdiği sonucuna ulaşılmıştır *(X^2 = 11.826; sd = 3; p = .008).*

Formlara ait madde faktör yüklerinin ikili karşılaştırmaları ise Wilcoxon İşaretli Sıralar Testi ile değerlendirilmiştir. Analiz sonucu, madde faktör yüklerinin ortancalarının Form 1 ile Form 3 ve Form 3 ile Form 4 arasında istatistiksel olarak anlamlı şekilde farklılaştığı (p<.05); diğer form kombinasyonları arasında ise manidar düzeyde bir farklılık olmadığı görülmüştür (p>.05). Cohen'in (1988) kriterlerine göre, madde faktör yüklerinin, Form 1 ile Form 3 arasında *(z = -2.825; p = .005; r = -.38)* ve Form 3 ile Form 4 arasında (*z = -2.661; p = .008; r = -.36)* orta etki büyüklüğü ile değişim gösterdiği belirlenmiştir.

Madde Faktör Yükleri Arasındaki İlişkiye Yönelik Bulgular

Üçüncü alt problemin çözümünde, formlardan hesaplanan madde faktör yükleri arasındaki ilişkinin ortaya çıkarılması için Spearman Korelasyon Katsayısı hesaplanmış ve sonuçlar Tablo 4'te sunulmuştur. Tablo 4'e göre, formlardan hesaplanan madde faktör yükleri arasında yüksek düzeyde ilişkiler olduğu görülmektedir. Tablo 4'e göre .95 korelasyon katsayısı ile madde faktör yükleri arasındaki en yüksek ilişki, Form 1 (orta noktada fikrim yok) ile Form 3 (orta noktada ne katılıyorum ne katılmıyorum) arasında ve Form 2 (orta değerde kararsızım) ile Form 3 arasında elde edilmiştir. En yüksek ikinci korelasyon katsayısı ise, Form 1 ile Form 2 arasında .90 şeklinde hesaplanmıştır. Bu bulguya dayanarak sözel ölçeklerden elde edilen madde faktör yüklerinin kendi aralarında oldukça yüksek düzeyde ilişki gösterdiği söylenebilir.

Tablo 4
Madde Faktör Yükleri Arasındaki İlişkiler

Formlar	Form 1	Form 2	Form 3	Form 4
Form 1	1.00			
Form 2	.90*	1.00		
Form 3	.95*	.95*	1.00	
Form 4	.80*	.83*	.85*	1.00

*p<.01

Tablo 4'te, ucu gömülü ölçek formundaki Form 4'ten elde edilen madde faktör yükleri ile sözel formlardan elde edilen madde faktör yükleri arasında .80-.85 aralığında yüksek ilişkiler olduğu görülmektedir. Form 4'ten elde edilen faktör yükleri ile en yüksek ilişkiyi Form 3 gösterirken, bunu takiben sırasıyla Form 2 ve Form 1 gelmektedir. Tablo 4'teki korelasyon katsayılarının tümü incelendiğinde ucu gömülü bir ölçek olan Form 4 ile sözel ölçekler arasında, sözel ölçeklerin kendi aralarındakinden daha düşük korelasyon katsayılarının hesaplandığı dikkat çekmektedir. Bir başka ifadeyle, sözel ölçeklerden elde edilen madde faktör yükleri

birbiriyle daha yüksek ilişkiye sahip iken; ucu gömülü ölçekten hesaplanan madde faktör yükleriyle daha düşük bir ilişki göstermiştir.

Formlardan Elde Edilen Toplam Puanlara İlişkin Bulgular

Dördüncü alt problemde, formlardan elde edilen toplam puanlar arasındaki farklılık tek yönlü varyans analizi ile test edildiğinde, form ortalamaları arasında istatistiksel olarak anlamlı bir farklılık olmadığı görülmüştür ($F_{3;1344} = 0.057$; $p > .05$). Bu bulguya dayanarak, orta değer etiketinin farklılaştığı sözel formlara ve ucu gömülü forma verilen öğrenci yanıtlarının ortalamalarının benzer olduğu söylenebilir. Bir başka deyişle, aynı amaca yönelik hazırlanmış, aynı ölçek maddelerine sahip ancak farklı kategori etiketlerine sahip dört farklı formdan hesaplanan öğrenci matematiksel düşünce (tutum) puanlarının ölçekler temelinde anlamlı bir farklılık göstermediği; öğrenci toplam puanlarının birbirine benzer olduğu söylenebilir. Bu bulgu, Dixon ve diğ. (1984), Finn (1972), Wyatt ve Meyers (1987), Newstead ve Arnold'ın (1989) bulguları ile benzerlik gösterirken; araştırmalarında çok boyutlu bir ölçek kullanan Jacko ve Huck (1974) ile çelişmektedir. Bu araştırmada kullanılan ölçek tek boyutlu bir ölçektir. Jacko ve Huck'ın (1974) çalışmasının bu araştırmadan farkı çalışmada çok boyutlu bir ölçeğin incelenmiş oluşudur. İki araştırma arasındaki bu uyuşmazlığın olası nedenlerinden birinin boyutluluk olabileceği düşünülmektedir.

Formlardan Elde Edilen Faktör Puanlarına İlişkin Bulgular

Beşinci alt problemde, formlardan elde edilen faktör puanları arasındaki farklılık tek yönlü varyans analizi ile test edildiğinde, faktör puanı ortalamaları arasında istatistiksel olarak anlamlı bir farklılık olmadığı görülmüştür ($F_{3;1344} = 0.000$; $p > .05$). Bu bulgudan hareketle, bir önceki bulguya benzer şekilde, orta değer etiketi farklı olan sözel formlardan ve ucu gömülü formdan hesaplanan faktör puanlarının ortalamaları benzerdir. Bir başka ifadeyle, öğrenci faktör puanları formlar arasında anlamlı bir farklılık göstermemiştir.

Model-Veri Uyumuna İlişkin Bulgular

Altıncı alt problemin çözümünde, formlara ait DFA sonuçları ve fark (Δ) değerleri aşağıda Tablo 5'te sunulmuştur. Tablo 5'te, χ^2/sd ve RMSEA değerleri incelendiğinde, tüm formlarda bu değerlerin düşük düzeyde uyuma işaret ettiği görülse de uyum indekslerinden TLI ve CFI değerleri tüm formlarda .95'in üzerinde hesaplanmıştır ki bu durum model-veri uyumunun çok iyi olduğunu göstermektedir. Ayrıca, SRMR .08'den küçük bir değere sahip olduğundan artıklar model uyumunun iyiliğini desteklemektedir. Bu bulgulara dayanarak, DFA sonuçlarının tüm formlarda iyi düzeyde model-veri uyumu sağladığı söylenebilir.

Tablo 5'te, form karşılaştırmalarının yapılabilmesi adında tüm formların ikili kombinasyonları üzerinden ΔRMSEA ve ΔCFI değerleri de sunulmuştur. Tablo 5'te yer alan ΔRMSEA değerlerinden yalnızca birinin $\pm.01$ ranjında olduğu; diğer fark değerlerinin ise bu sınırı aştığı dikkat çekmektedir. Chen'e (2007) göre, ΔRMSEA .01'den büyük olduğu zaman farklılık önemlidir. Buna göre, model-veri uyumu

açısından, Form 2 ve Form 3 arasında önemli bir farklılık olmadığı; diğer formlar arasındaki farklılıkların ise önemli olduğu yorumu yapılabilir. Tablo 5'e göre, Form 1'in diğer formlara kıyasla veriye daha iyi uyum sağladığı söylenebilir. DFA sonucu elde edilen madde faktör yükleri Ek'te yer almaktadır.

Tablo 5
Formlara Ait DFA Sonuçları ve Fark Değerleri

Formlar	χ^2	χ^2/sd	RMSEA	TLI	CFI	SRMR	ΔCFI	ΔRMSEA
Form 1	306.66	3.98	.09	.99	.99	.04		
Form 2	416.51	5.41	.12	.99	.99	.05		
Form 3	437.63	5.68	.12	.98	.99	.05		
Form 4	544.35	7.07	.13	.99	.99	.06		
	Form 1 ve Form 2 arasındaki fark değerleri						.00	-.02
	Form 1 ve Form 3 arasındaki fark değerleri						.00	-.02
	Form 1 ve Form 4 arasındaki fark değerleri						.00	-.04
	Form 2 ve Form 3 arasındaki fark değerleri						.00	.00
	Form 2 ve Form 4 arasındaki fark değerleri						.00	-.02
	Form 3 ve Form 4 arasındaki fark değerleri						.00	-.02

Ölçme Değişmezliğine Yönelik Bulgular

Yedinci alt problemin çözümünde çoklu grup DFA işe koşulmuştur. Form 1, Form 2, Form 3 ve Form 4'ün ölçme değişmezliğini ne düzeyde sağladığına ilişkin bulgular aşağıda Tablo 6'da sunulmuştur. Alanyazında. hesaplanan Ki-kare değerlerinin farkı (ΔX^2) manidar değilse ve CFI, TLI, RMSEA fark değerleri ± .01 aralığında ise ilgili değişmezlik türünün sağlandığı kabul edilmektedir (Kline, 2016).

Tablo 6'daki ΔX^2 değerlerine ait anlamlılık düzeylerinin metrik ve skalar değişmezliğin sağlanabileceğini ($p>.05$); ancak katı değişmezliğin sağlanamayacağını göstermektedir ($p<.05$). ΔCFI ve ΔRMSEA değerleri incelendiğinde ise, bu değerlerin ± .01 ranjını aşması sebebiyle, metrik değişmezliğin sağlanamadığı görülmektedir. Metrik değişmezliğin sağlanması durumunda skalar değişmezlik test edilmektedir (Gregorich, 2006). Bu noktada metrik değişmezliğin sağlanamaması sebebiyle skalar değişmezlikten elde edilen sonuçlar yorumlanmamıştır. Bu sonuca dayanarak formların metrik değişmezliği sağlamadığı ve sadece yapısal değişmezliğe sahip olduğu söylenebilir.

Tablo 6
Ölçme Değişmezliği Sonuçları

Değişmezlik	χ^2	sd	$\Delta\chi^2$	p	CFI	RMSEA	TLI	ΔCFI	ΔRMSEA	ΔTLI
Yapısal	328.45	308	-	-	.92	.10	.90	-	-	-
Metrik	422.30	347	37.57	.54	.98	.05	.98	.06	-.05	.08
Skalar	440.48	386	47.74	.16	.98	.05	.98	-.00	-.00	.00
Katı	487.05	428	60.59	.03	.98	.04	.98	.00	-.00	.00

Not. CFI=Comparative Fit Index; TLI=Tucker-Lewis Index; RMSEA=Root Mean Square Error of Approximation

Tartışma, Sonuç ve Öneriler

Bu araştırmada, aynı amaca yönelik hazırlanmış aynı maddelerden oluşan dört farklı türdeki Likert tipi ölçek uygulaması üzerinden ölçeklerin geçerlik ve güvenirlik kanıtlarının karşılaştırılması amaçlanmıştır. Bu amaçla ucu gömülü bir form (Form 4) ve farklı orta değer etiketine sahip üç adet sözel form (Form 1, Form 2, Form 3) ele alınmıştır. Sözel formların orta değer etiketleri için Form 1'de "Fikrim yok" ifadesi; Form 2'de "Kararsızım" ifadesi ve Form 3'te "Ne katılıyorum ne katılmıyorum" ifadesine yer verilmiştir. Form 4'te ise ilk ve son değerleri sözel etiketle gösterilen ve geriye kalan kısmı sayısal değerlerden oluşan ucu gömülü bir form kullanılmıştır. Araştırmanın odak noktasını, Likert tipi ölçeklerde farklı orta değer etiketlerinin ve ölçek versiyonlarının arasındaki ilişkinin ortaya çıkarılması ve konuya ilişkin ampirik bir kanıt sunulması oluşturmaktadır.

Araştırma kapsamında uygulanan sözel ve ucu gömülü formların toplam puan ortalamaları ve faktör puanları arasında manidar bir farklılık olmadığı sonucuna ulaşılmıştır. Bu bulgu, Dixon ve diğ. (1984), Finn (1972), Wyatt ve Meyers (1987), Newstead ve Arnold (1989) bulguları ile örtüşmektedir. Dixon ve diğ. (1984), bu çalışmadan farklı olarak, çalışmalarında altı kategorili ve dolayısıyla zorunlu cevaplı (forced choice) bir puanlama kullanmış; sözel ve ucu gömülü formlar arasında ortalamalar bakımından bir farklılık olmadığını ortaya koymuştur. Benzer şekilde, Finn (1972), Wyatt ve Meyers (1987), Newstead ve Arnold (1989) da, Likert tipi ölçeklerin farklı etiket formatlarını karşılaştırdığı çalışmalarında, etiket tanımlarının puanlar üzerinde anlamlı bir farklılık ortaya çıkarmadığını göstermiştir. Bu çalışmada beşli Likert puanlaması üzerinden karşılaştırmalar gerçekleştirilmiş ve ortalamalar bakımından benzer sonuçlara ulaşılmıştır. Bu çalışmanın bulgularının aksine, Jacko ve Huck (1974), çok boyutlu bir ölçek olan Alpert-Haber Başarı Anksiyetesi Testi (Alpert-Haber Achievement Anxiety Test; AHAAT) üzerinden, ucu gömülü ölçeklerin sözel ölçeklere kıyasla daha düşük ortalamaya sahip olduğunu belirtmiştir. Bu çalışmada veri toplama aracı olarak kullanılan Matematiğe İlişkin Düşünceler Ölçeği (MİDÖ) araştırmanın yöntem bölümünde de belirtildiği gibi tek boyutlu bir ölçektir. Bu bağlamda Jacko ve Huck'un (1974) belirttiği gibi, AHAAT ölçeğinin tek boyutlu olmaması sonuçların uyuşmazlığının bir sebebi olarak düşünülebilir. Bu

çalışmanın sonuçlarına göre, aynı özelliği ölçmek üzere aynı maddelerden oluşan kategori etiketleri farklı dört formun bireyler hakkında benzer bilgiler sundukları görülmüştür. Newstead ve Arnold'ın (1989) belirttiği üzere, formlar arasında toplam puan bakımından manidar bir farklılık, farklı etiketlerin kullanılması durumunda ölçek puanlarının artması veya azalması anlamına gelecektir ki bu durum farklı formatlardan elde edilen puanların karşılaştırılamayacağına işaret eder. Bu açıdan dört formun karşılaştırılabilir olduğunu söylemek uygun olacaktır. İleride yapılacak çalışmalar için araştırmacılara, çok boyutlu ve zorunlu cevaplı ölçekler üzerinden de benzer çalışmaların tekrarlanması önerilebilir.

Bu araştırmadan elde edilen ikinci sonuç, formların tümünün polikorik korelasyon matrisine dayalı AFA sonucu açıklanan varyans oranlarının ve güvenirliğinin birbirine yakınlığıdır. Ölçek orta noktası "Ne katılıyorum ne katılmıyorum" şeklinde etiketlendiğinde (Form 3) açıklanan varyans en yüksek değeri alırken; bunu takiben sırasıyla "Kararsızım" etiketli Form 2, "Fikrim yok" etiketli Form 1 ve ucu gömülü Form 4 gelmektedir. Form 1 ve Form 4'ün oldukça yakın açıklanan varyans oranına sahip olması bu iki formun benzer düzeyde bilgi verdiği şeklinde yorumlanabilir. Formların tümüne ait Cronbach Alfa iç tutarlılık katsayısı .97 ve .98 olarak hesaplanmıştır. Buna göre formların güvenirliklerinin benzer ve yüksek düzeyde olduğu yorumu yapılabilir. Bu bulgu, Finn (1972), Wyatt ve Meyers (1987) ile Weng'in (2004) çalışmasındaki bulgular ile benzerlik gösterirken; Jacko ve Huck (1974) ile Krosnick ve Berent'in (1993) çalışmaları ile örtüşmemektedir. Az önce de belirtildiği gibi Jacko ve Huck (1974) araştırmasında çok boyutlu bir ölçekten yararlanmıştır. Krosnick ve Berent (1993) ise çalışmasında, bu araştırmadan farklı olarak, katılımcıların politik görüşlerini belirlemek üzere yedili Likert tipinde bir ölçek kullanmış ve bu ölçeği, telefonla, yüz yüze ve cevaplayıcının kendi yanıtlarını kendi kaydettiği şekillerde farklı oturumlarda uygulamış ve uygulamalar arasındaki uyumu güvenirlik olarak ele almıştır. Bu açıdan çalışma sonuçlarının örtüşmemesi söz konusu olmuş olabilir. Buna dayanarak konu hakkında ileride yapılacak çalışmalara farklı uygulamalar (örneğin yüz yüze ve çevrimiçi) arasındaki uyumun incelenmesi, bu açıdan AFA ve DFA sonuçlarının karşılaştırılması önerilebilir.

Bu araştırma sonucunda, AFA sonucu elde edilen madde faktör yüklerinin, orta değerde "Fikrim yok" ifadesi ile "Ne katılıyorum ne katılmıyorum" ifadelerinin yer alması durumunda ve ölçeğin ucu gömülü sunulması ile orta değerde "Ne katılıyorum ne katılmıyorum" ifadesinin yer alması durumunda orta etki büyüklüğü ile anlamlı bir farklılık ortaya çıkardığı görülmüştür. Bununla birlikte, formlardan elde edilen madde faktör yükleri arasındaki korelasyonlara göre, sözel ölçeklerden elde edilen madde faktör yüklerinin kendi aralarında daha yüksek bir ilişkiye sahip olduğu; fakat ucu gömülü ölçekten elde edilen madde faktör yükleriyle daha düşük bir ilişki gösterdiği sonucuna ulaşılmıştır. Bir diğer anlatımla, sözel tutum etiketlerinden oluşan formların madde faktör yükleri arasında daha yüksek düzeyde ilişki görüldüğü ve sözel formların ucu gömülü form ile daha düşük ilişkiye sahip olduğu söylenebilir. Buna göre ölçeklerin tutum etiketlerindeki sayısal-sözel ifade farkı, ölçülecek özelliğe ilişkin bireylerin tepkisini değiştiriyor yorumu yapılabilir. Bu sonuca dayanarak

bireylerin aynı kategoride yer alan farklı tutum etiketlerini farklı şekillerde algıladıkları söylenebilir. Araştırmacılara, benzer araştırmaların farklı özellikleri ölçen ölçekler üzerinden gerçekleştirilmesi önerilebilir.

Araştırmanın ilginç bir sonucu olarak, ΔRMSEA değerlerine göre yalnızca bir karşılaştırmada sınırın aşılmadığı dikkat çekmektedir. Orta değerde "Kararsızım" ifadesinin yer aldığı Form 2 ile orta değerde "Ne katılıyorum ne katılmıyorum" etiketinin yer aldığı Form 3 arasında model-veri uyumu bakımından önemli bir farklılık olmadığını söylemek mümkündür. Model-veri uyumunun karşılaştırıldığı diğer form kombinasyonlarında ise ΔRMSEA değerlerine göre önemli farklılıklar olduğu ve formlar arasından veriye daha iyi uyum sağlayan formun orta değeri "Fikrim yok" etiketli olan Form 1 olduğu görülmüştür. Buna dayanarak, orta değerde "Fikrim yok" ifadesinin bulunması, diğer sözel ölçeklere ve ucu gömülü ölçeğe kıyasla öğrencilerin tepkisi için daha uygundur diyebiliriz. Oysa Başar'a (2021) göre, "kararsızım, fikrim yok, bir şey söyleyemem" gibi ifadeler orta derece bir tutum düzeyi değil; durum bildirmektedir. Fakat Nadler ve diğ.'ne (2015) göre tutum ifadelerinde orta derece "nötr" anlamına gelmesinin yanında, formu dolduran bireylerin merkeze yönelme yanlılığını ve sosyal beğenilme arzusunu da içermektedir. Bu nedenle Form 1'in model-veri uyumu daha iyi çıkmış olabilir.

Ölçme değişmezliği bakımından elde edilen sonuca göre, aynı özelliği ölçen, aynı ölçek maddelerine sahip fakat farklı kategori etiketlerine sahip formlar arasında yalnızca yapısal değişmezliğin sağlandığı; metrik değişmezliğin ise sağlanamadığı ortaya konulmuştur. Formlar arasında metrik değişmezliğin sağlanamaması, madde faktör yüklerinin formlarda farklılık gösterdiğine işaret etmektedir. Bu durumda maddelerde formlara göre madde yanlılığı (DIF) ya da madde parametresi kayması (DRIFT) olup olmadığı nicel ve/veya nitel yöntemlerle incelenmesi önerilebilir. Bu araştırmanın bulguları ışığında ileride yapılacak çalışmalar için araştırmacılara, bu tür araştırmaların sınırlı sayıda olması sebebiyle tekrarlanması, sözel, ucu gömülü ve ayrıca sayısal formların geçerlik ve güvenirlik bakımından incelenmesi, ölçek maddelerinin faktör yüklerinin farklı tekniklerle karşılaştırılması ile değişkenlik kaynaklarının Genellenebilirlik Kuramı ve Madde Tepki Kuramı temelinde belirlenmesi önerilebilir.

Covid-19 pandemisi sebebiyle bu araştırmada veri toplama aşaması çevrimiçi bir platformda gerçekleşmiştir. Verinin toplanması aşamasında formların linkleri sırasıyla Form 1, Form 2, Form 3 ve Form 4 olacak şekilde birer hafta arayla dört gün boyunca öğrencilerin erişimine açık bırakılmıştır. Bu süreci biraz daha açıklamak gerekirse öğrencilere öncelikle Form 1 uygulanmıştır. Bu formun linki dört gün boyunca öğrenci erişimine açık bırakılmış ve dört günün sonunda erişime kapatılmıştır. Form 1'e erişimin kapanmasının ardından tam bir hafta sonra Form 2'nin linki öğrencilere dağıtılmış ve formların tümü için bu zamanlamaya özen gösterilmiştir. Formların farklı öncelik ve sonralık kombinasyonları dikkate alınarak veri toplamanın pandemide zor olacağı ve öğrencilere karışık gelebileceği düşünülerek form uygulama sırası bu araştırmada sabit tutulmuştur. Bu durumda sıra

etkisinden kaynaklanan hataların ölçme sürecini etkileyebileceği düşünülebilir. Bu durum araştırmanın bir sınırlılığı olarak karşımıza çıkmaktadır. İleride yapılacak araştırmalarda sonuçları sıra etkisinden arındırmak amacıyla formların farklı öncelik-sonralık kombinasyonlarına yer verilmesi önerilebilir. Bununla birlikte bu araştırmada temel alınan ölçek beşli Likert tipinde puanlanan, 14 madde ve tek boyuttan oluşan MİDÖ'dür. Daha fazla ya da az sayıda madde içeren, üçlü veya dörtlü Likert gibi farklı türde puanlanan ve çok boyutluluk özelliği gösteren ölçekler üzerinden bu tip araştırmaların tekrarlanması önerilebilir. Ayrıca bu araştırmada sözel ölçeklerde yalnızca orta değer etiketleri değiştirilmiş; diğer dört etiket aynı bırakılmıştır. Farklı etiket ifadelerinin tutum üzerindeki etkisi tüm etiket ifadeleri değiştirilerek veya etiketler farklı şekillerde sıralanarak da incelebilir.

References

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological Methodology, 22,* 83-118. doi:10.2307/270993

Anastasi, A. (1982). *Psychological testing* (5th ed.). Macmillian.

Balcı, A. (2011). *Sosyal bilimlerde araştırma: Yöntem, teknik ve ilkeler [Research in social sciences: Methods, techniques and principles]* (9. baskı). Pegem Akademi.

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9*(2), 151-160. doi: 10.1207/s15324818ame0902_4

Başar. H. (2021). *Araştırmalarda Likert yanılgıları*. *[Likert misconceptions in research]*. http://docplayer.biz.tr/52855457 adresinden 07.10.2021 tarihinde erişilmiştir.

Bendig, A. W. (1954). Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology, 38(*3), 167–170. doi: 10.1037/h0059072

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

Bora Semiz. B., & Altunışık R. (2016). Pazarlama araştırmalarında Likert tipi ölçeklerin özelliklerinin cevaplama tarzları üzerindeki etkilerinin incelenmesi [An evaluation of various attributes of Likert type scales on response styles in marketing research]. *Bartın Üniversitesi İİBF Dergisi, 7*(14), 577-598. https://dergipark.org.tr/tr/download/article-file/251735

Bryne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross Cultural Psychology, 34*(2), 155-175. doi: 10.1177/0022022102250225

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504. doi: 10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. doi: 10.1207/S15328007SEM0902_5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

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

Ferrando, P. J., & Lorenzo-Seva U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*, 762-780. doi:10.1177/0013164417719308

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

Gegez. E. (2010). *Pazarlama araştırmaları [Marketing studies]* (3. baskı). Beta Yayınları.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*, 78-94. doi: 10.1097/01.mlr.0000245454.12228.8f

Hartley, S. L., & MacLean, W. E. (2006). A review of the reliability and validity of Likert-type scales for people with intellectual disability. *Journal of Intellectual Disability Research, 50*(11), 813-827. doi: 10.1111/j.1365-2788.2006.00844.x

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

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1217–1218. doi: 10.1111/j.1365-2929.2004.02012.x

Kağıtçıbaşı. Ç. (2010). *Günümüzde insan ve insanlar [People and people today]* (14. baskı). Evrim Yayınevi.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34*(1), 111 -117. doi: 10.1177/001316447403400

Kılıç, A. F., Uysal, İ., & Kalkan, B. (2021). An alternative to Likert scale: Emoji. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 182-191. doi: 10.21031/epod.864336

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

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

Krosnick. J. A., & Presser. S. (2010). Questionnaire design. J. D. Wright & P. V. Marsden (Eds.). in *Handbook of Survey Research* (pp. 263-313). Emerald Group.

Kurtuluş. K. (2006). *Pazarlama araştırmaları [Marketing studies] (*8.baskı). Literatür Yayıncılık.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing and Health, 25*, 295–306. doi: 10.1002/nur.10041

Lorenzo-Seva, U., & Ferrando, P. J. (2021). Factor (Version 11.05.01) [Computer software]. Tarragona: Universitat Rovira i Virgili.

Mardia, K. V. (1970). Measures of multivariate skewnees and kurtosis with applications. *Biometrika, 57*, 519-530. doi:10.2307/2334770

Nadler. J. T., Weston. R., & Voyles. E. C. (2015). Stuck in the middle: The use and interpretation of mid points in items on questionnaires. *The Journal of General Psychology. 14*(2), 71-89. doi: 10.1080/00221309.2014.994590

Nakip. M. (2006). *Pazarlama araştırmaları: Teknikler ve (spss destekli) uygulamalar [Marketing studies: Techniques and applications with spss]*. Seçkin Yayıncılık.

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik Likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi [The investigation of item and scale properties of Likert type scale and metric scale measuring the same attitude according to classical test theory and item response theory]* (Tez No. 113510) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Başkanlığı Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

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

O'Muircheartaigh, C. A., Krosnick, J. A., & Helic, A. (2000). Middle alternatives, acquiescence, and the quality of questionnaire data. Working Papers 0103, Harris School of Public Policy Studies, University of Chicago.

Payne, S. L. (1950). Thoughts about meaningless questions. *Public Opinion Quarterly, 14*, 687–696. doi: 10.1086/266248

Peters, D. L., & McCormick, E. J. (1966). Comparative reliability of numerically anchored versus job -task anchored rating scales. *Journal of Applied Psychology, 50,* 92-96. doi: 10.1037/h0022935

R Core Team. (2013). *R: A language and environment for statistical computing*, (Version 3.0.1), Vienna, Austria: R Foundation for Statistical Computing. Online: http://www.R-project.org/

Robie, C., Meade, A. W., Risavy, S. D., & Rasheed, S. (2022). Effects of response option order on Likert type psychometric properties and reactions. *Educational and Psychological Measurement, 82*(6), 1107–1129. doi: 10.1177/00131644211069406

Seashore, S. E., & Katz, D. (1982). Obituary: Rensis Likert (1903-1981). *American Psychologist, 37*(7), 851-853. doi: 10.1037/0003-066X.37.7.851

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*, 557–566. doi: 10.1037/pas0000648

Tezbaşaran. A. (2008). *Likert tipi ölçek geliştirme kılavuzu [Likert type scale development guide]* (3. baskı). Türk Psikologlar Derneği Yayınları.

Turgut, M. F., & Baykul, Y. (1992). *Ölçekleme teknikleri [Scaling techniques]*. ÖSYM Yayını.

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*, 956–972. doi: 10.1177/0013164404268674

Wu, D. A., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1-26. doi: 10.7275/mhqa-cd89

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

Vaillancourt, P. M. (1973). Stability of children's survey responses. P*ublic Opinion Quarterly, 37*, 373-387. doi: 10.1086/268099

Yükselen. C. (2003). *Pazarlama araştırmaları [Marketing studies]* (2. baskı). Detay Yayıncılık.

Ethical Declaration and Committee Approval

In this research, the principles of scientific research and publication ethics were followed.

Ethics committee approval was received for this study from Inonu University Scientific Research and Ethics Committee (Date: 02/07/2021, No: 13-19).

Bu çalışma için etik komite onayı, İnönü Üniversitesi Bilimsel Araştırmalar ve Etik Kurulu'ndan (Tarih: 02/07/2021, No: 13-19) alınmıştır.

Proportion of Author's Contribution

Each of the authors contributed equally to the article.

Appendix

Table 7

Item Factor Loads calculated from CFA

Item	Form 1	Form 2	Form 3	Form 4
M1	0.93	0.93	0.94	0.92
M2	0.80	0.83	0.84	0.87
M3	0.94	0.95	0.96	0.96
M4	0.92	0.95	0.94	0.92
M5	0.89	0.86	0.92	0.88
M6	0.88	0.91	0.89	0.87
M7	0.91	0.95	0.94	0.88
M8	0.70	0.68	0.71	0.68
M9	0.87	0.90	0.91	0.92
M10	0.90	0.92	0.90	0.92
M11	0.80	0.84	0.84	0.81
M12	0.92	0.92	0.94	0.90
M13	0.82	0.85	0.84	0.86
M14	0.95	0.94	0.94	0.89