International Journal of Assessment Tools in Education
2025, Vol. 12, No. 2, 341–352
https://doi.org/10.21449/ijate.1581217
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Exploring number of response categories in factor analysis: Implications for sample size

Fatih Orçan[1*]

[1]Kahramanmaraş Sütçü İmam University, Faculty of Education, Department of Educational Sciences, Türkiye

ARTICLE HISTORY
Received: Nov. 8, 2024
Accepted: Jan. 30, 2025

Keywords:
Factor analysis,
Sample size,
Number of response categories,
Number of items.

Abstract: Factor analysis is a statistical method to explore the relationships among observed variables and identify latent structures. It is crucial in scale development and validity analysis. Key factors affecting the accuracy of factor analysis results include the type of data, sample size, and the number of response categories. While some studies suggest that reliability improves with more response categories, others find no significant relationship between the number of response categories and reliability. A key consideration is that increasing the number of response categories can introduce measurement errors, especially when there are too many categories for participants to respond accurately. The study examines how different numbers of response categories affect sample size requirements in factor analysis, particularly under misspecified and correctly specified models. MonteCarloSEM package in R was used to simulate data sets based on sample size, number of response categories, model specification, and test length. Results show that a higher number of categories helps reduce bias and improve model fit, especially in smaller samples. However, when sample sizes are small or when fewer categories are used, increasing the number of items or the number of categories can improve parameter estimation. The findings suggest that for optimal results, researchers should carefully balance sample size, number of items, and response categories, particularly in studies with categorical data.

1. INTRODUCTION

Factor analysis is a multivariate statistical technique that aims to explore the relationships among observed variables and uncover the underlying latent structures of these variables. This technique, which is widely used in the fields of social sciences and psychometrics, plays a crucial role in scale development and validity analysis. The type of data used in factor analysis and the sample size are significant factors affecting the accuracy of the results obtained (Kyriazos, 2018). In factor analysis, the number of response categories used in items can influence the quality and reliability of the results (Bandalos & Enders, 1996; Lozano *et al*., 2008). While researchers often propose different recommendations, there is no clear consensus on the best number of response categories (Simms *et al*., 2019; Wakita *et al*., 2012; Yoon, 2024). Abulela and Khalaf (2024) note that while some studies recommend between 4 and 7

*CONTACT: Fatih ORÇAN ✉ fatihorcan@ksu.edu.tr Kahramanmaraş Sütçü İmam University, Faculty of Education, Department of Educational Sciences, Kahramanmaraş, Türkiye

e-ISSN: 2148-7456

response categories, others focus on the range between 5 and 7, however; most scales use 4 to 5 categories.

Several studies investigate the relationship between the number of response categories and the reliability of measurement (Abulela & Khalaf, 2024; Lozano *et al*., 2008; Matell & Jacoby, 1971; Wakita *et al*., 2012; Yoon, 2024). While many argue that reliability improves with more categories, some studies contradict this statement. That is, some researchers suggest that increasing the number of categories improves reliability, while other studies show no significant relationship between reliability and the number of response categories (Abulela & Khalaf, 2024; Yoon, 2024). Specifically, "both reliability and validity are independent of the number of response categories" (Matell & Jacoby, 1971). The number of response categories does not significantly affect descriptive statistics or Cronbach's alpha (Wakita *et al*., 2012). Having more categories does not necessarily guarantee higher reliability (Abulela & Khalaf, 2024) nor does reducing longer scales to two-point or three-point scales may compromise the reliability or validity of the results (Matell & Jacoby, 1971).

Increasing the number of categories can also lead to drawbacks. Komorita and Graham (1965, as cited in Abulela & Khalaf, 2024) suggest that using more than seven categories may exceed participants' capacity to respond accurately, which increases measurement error. The complexity of a 7-point scale is unwarranted; four- or five-point scales provide adequate differentiation for most measurement needs (Wakita *et al*., 2012). Participants tend to avoid selecting extreme responses on a 7-point scale, as "an increase in the number of options biases responders against answering the strongest expressions" (Wakita *et al*., 2012, p. 543). For instance, Abulela and Khalaf (2024) point out that using more than seven categories can make labeling each option more challenging. Similarly, Simms *et al*. (2019) argues that there is no psychometric advantage (e.g., in terms of alpha) to using more than six categories. In fact, presenting respondents with many similar options (e.g., "strongly disagree," "disagree," "slightly disagree") can make it more difficult to differentiate between them, potentially leading to confusion.

Increasing the number of categories can also increase the time required for respondents to complete the questionnaire (Preston & Colman, 2000). According to Preston and Colman (2000), 5-point scales are the easiest to use, but they may be insufficient for expressing emotions or thoughts compared to scales with more categories. However, items with fewer categories are faster to answer compared to those with more categories, and thus scales with as few as 3 categories can be used depending on the study's purpose (Preston & Colman, 2000). Additionally, the number of categories may depend on the respondents' age group. For example, 3 to 4 categories may be appropriate for children, while adults may benefit from scales with 5 or more categories (Abulela & Khalaf, 2024). This is similar to the increase in response options for multiple-choice questions that accompany an increase in test-takers' age. Therefore, if participants find it too difficult or too easy to express their attitudes or feelings using the scale, their motivation to respond may decrease, leading to lower-quality data (Preston & Colman, 2000). Moreover, if the number of response categories is not chosen carefully when drafting items, it may result in significant measurement errors during parameter estimation (Abulela & Khalaf, 2024; Bandalos & Enders, 1996).

From another point of view, increasing the number of response categories leads to better model fit, regardless of whether the model is correctly specified or misspecified (Maydeu-Olivares *et al*., 2017). It was suggested that scales with more than 5 categories may help detect misspecified models more effectively. However, it was also pointed out that reducing the number of response alternatives may also decrease the likelihood of rejecting a misspecified model (Maydeu-Olivares *et al*., 2017). Specifically, decreasing the number of categories can improve model fit while introducing parameter bias, thus reducing the likelihood of rejecting a model that should be rejected (Abdelsamea, 2020). Moreover, the Root Mean Square Error of Approximation

(RMSEA) tends to increase with the number of response categories, whereas the impact of the number of categories on SRMR is less pronounced (Maydeu-Olivares *et al*., 2017).

The sample size is also a critical factor in factor analysis regarding model fit and the accuracy of parameter estimates (Kline, 2011). In particular, the sample size plays a crucial role in the effectiveness of factor analysis (MacCallum *et al*., 1999). Rhemtulla *et al*. (2012) defined studies with sample sizes ranging from 100 to 200 as small, and those with up to 600 participants were categorized as medium-sized. However, researchers often debate the ideal sample size, as this can vary depending on factors such as the complexity of the model, the type of data, and the analytical approach used. Kyriazos (2018) provides a comprehensive overview of the various factors that influence sample size in factor analysis. Specifically,

> "In CFA, being a SEM category, sample size depends on a number of features like study design (e.g. cross-sectional vs. longitudinal); the number of relationships among indicators; indicator reliability, the data scaling (e.g., categorical versus continuous) and the estimator type (e.g., ML, robust ML etc.), the missing data level and pattern and model complexity" (p. 2208).

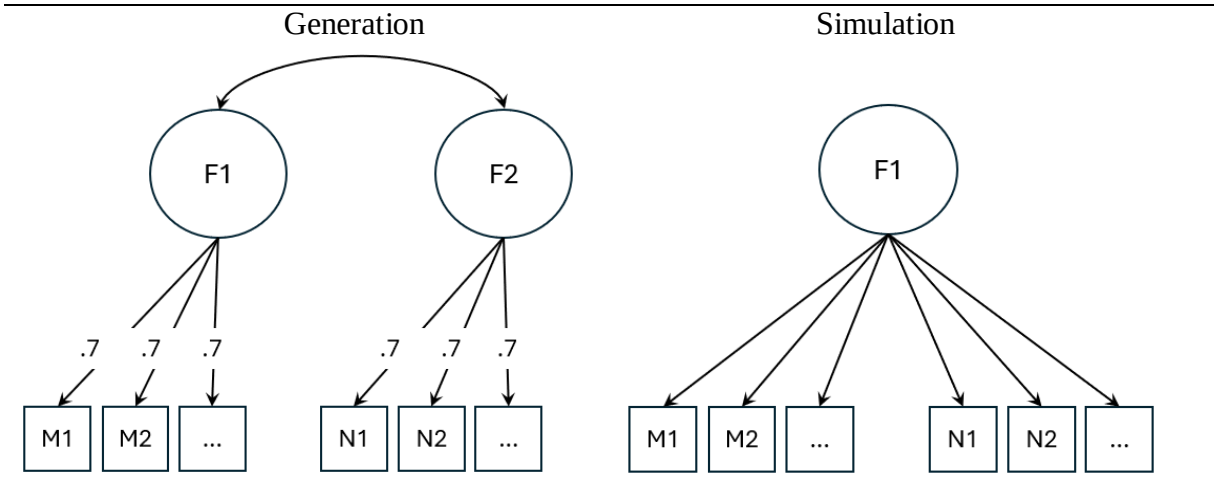
While the distinction between categorical and continuous data is often emphasized in discussions of sample size, the number of categories within categorical data also plays a critical role. The relationship between the number of categories and sample size remains underexplored in the literature. The purpose of this study is to examine the effects of the number of response categories in scales on sample size requirements in factor analysis.

2. METHOD

2.1. Data Generation Procedure

The data generation process was conducted using the MonteCarloSEM package (Orçan, 2021) within the R-CRAN environment (R Core Team, 2020). This package facilitates the simulation and analysis of data sets under various simulation conditions, including sample size and normality, for a specified model (Orçan, 2021). The data were generated based on two-factor models as shown in Figure 1. Based on the simulation conditions, the data generation model was modified. Various design factors were considered in this simulation study. After the data were generated, a one-factor model, as shown in Figure (right panel), was used to analyze the data.

Figure 1. *Models for data generation and simulation.*



2.1.1. *Design factors for data generation*

Four design factors were considered for the study.

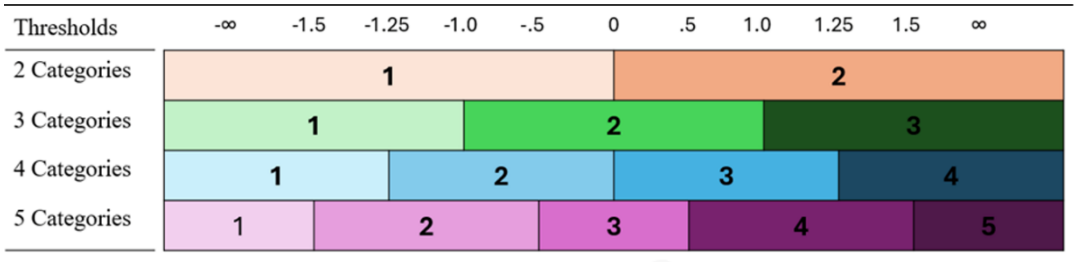
- Sample Sizes: Five different sample sizes were considered: 100, 200, 300, 500 and 1000. These sample sizes were deliberately chosen to cover a broad spectrum, ranging from small to large, within the context of Structural Equation Modeling (SEM) analysis (Orcan & Yang, 2016).
- Test Length: Three different test lengths were considered: 6, 10, 20. The number of items was equally distributed across factors. For example, when there are 6 items in total, three of those were assigned to the first factor and the rest assigned to the second factor. The factor loadings were all set at .7 within the study. That is, the factor loading was not a design factor for this study.
- Number of Response Categories: Four different categories were used within this study: 2 (C2), 3 (C3), 4 (C4), and 5 (C5). Items with two categories represent responses such as yes/no or pass/fail. Similarly, five categories represent five-point Likert-type items such as strongly disagree, disagree, neutral, agree, strongly agree. Categories larger than five were not considered in this study since scales with four or five points are sufficient (Wakita *et al*., 2012).
- Model Specifications: Three different models were used in this study: Strong misspecified model (r = .5), moderate misspecified model (r = .8), and correctly specified model (r = 1). Under the strongly misspecified model (MS_S), the correlation between the factors was set to be .5. However, the data analyzed with one-factor model as if the correlation between the factors was assumed to be 1. Under the moderately misspecified model (MS_M), the correlation was .8 and in the correctly specified model, the correlation was set to 1. That is, under the correctly specified (CS) model, the data generation and simulation model were the same.

Consequently, the data were generated based on a total of 180 distinct conditions using these four design factors: 5 sample sizes x 3 Number of items x 4 number of categories x 3 model specifications.

2.2. Data Analysis

A total of 1000 data sets were generated using the MonteCarloSEM package for each of the conditions. The data generation process began by generating normally distributed data sets. Later, using per-given threshold values, the normally distributed data sets were transformed into categorical data sets. The threshold values which were used are given in Figure 2. For example, to create two categories, the threshold value was set to 0 (zero). The simulated values lower than 0 were set to 1 and larger than 0 were set to 2 to create two categories (Kılıç, 2022). The CFA modes were estimated using the Weighted Least Squares with Mean and Variance (WLSMV) estimation method in the lavaan packages (Rosseel, 2012). The WLSMV is considered "the best available categorical estimator" and is recommended for data sets with variables containing fewer than five categories (Rhemtulla *et al*., 2012, p. 354). Moreover, the WLSMV method could produce model solutions with sample sizes as small as 100 (Flora & Curran, 2004), indicating its robustness even in smaller datasets.

Figure 2. *Threshold values used for categorical data generation.*



Model data fits were estimated using the p-value from the chi-square test. In addition to the chi-square test, supplementary fit indices such as the Comparative Fit Index (CFI) and RMSEA should be examined to obtain a comprehensive evaluation of model fit (Flora & Curran, 2004). Therefore, CFI, RMSEA, and the standardized root mean square residual (SRMR) values were also assessed for model-data fit. For the evaluation of model data fit, Hu and Bentler's (1999) criteria were used. Also, parameter estimates for the first factor loadings were examined in detail. For this purpose, the relative biases were calculated using Equation 1.

$$Bias = \frac{Abs(Estimated\ Value - True\ Value)}{True\ Value} \quad (1)$$

Although Flora and Curran (2004) indicated that biases below 5% are considered "trivial", the critical bias threshold was set at 5% in order to be more conservative. The factor loadings for the models were all set at .7 within the study. Therefore, the true value in this study was .7.

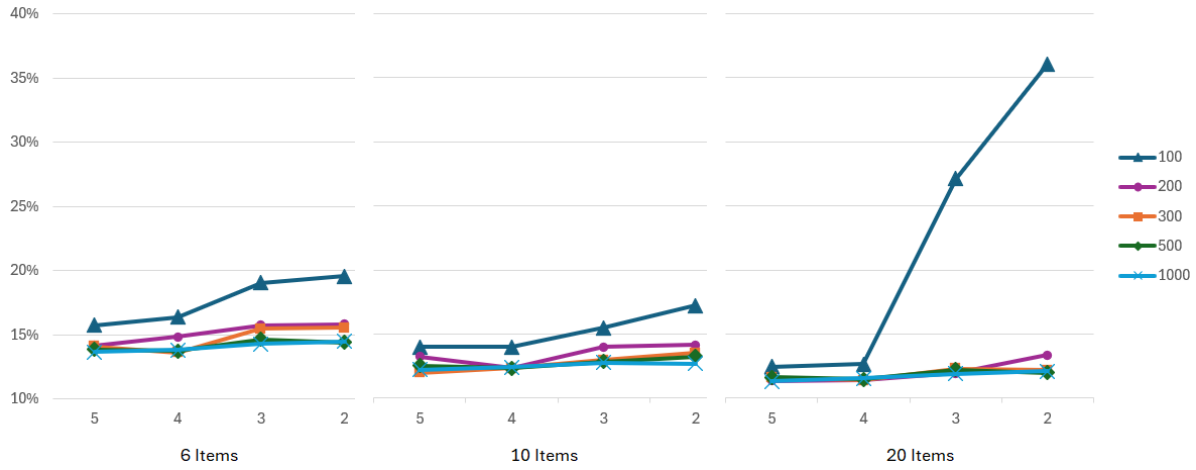
3. RESULTS

All the replications across the models converged to a solution. That is, non-converge was not an issue for the simulation conditions.

3.1. Results Based on Mis-specified Models

Two misspecified models were examined in this study. Results based on the bias calculations are presented in Figure 3 and Figure 4. When the correlation between the factors was .5, all the bias estimates exceeded the 5% critical value (see Figure 3). When examined in detail, as the sample size increased, the percentage of bias decreased, regardless of test length. Also, when the number of items increased, the percentage of bias decreased slightly. However, the decreased bias values were still much higher than the 5% critical value.

Figure 3. *Bias of the parameter estimates for MS_S.*



In short, under the MS_S (larger miss specification), having fewer categories for an item does not require more items or larger sample sizes, except when the sample size is too small, such as 100. That is, when the sample size is larger than 100, the estimated bases were almost identical, regardless of the number of items or categories. However, when the sample size was 100, confusing results were obtained.

When the misspecification level was decreased, i.e., when the correlation between the factors was set to .8 (MS_M), the levels of bias decreased for all conditions. Although the bias values decreased, similar patterns can be observed across the figures. Specifically, as the sample size increased, the percentage of bias decreased. Under the small sample size conditions, the increasing number of items affected the percentage of bias. For instance, when the sample size was 200 and the number of categories was 2, the bias was 10% for 6 items; however, increasing the number of items to 10 reduced the bias to 8%. Under small sample sizes, it appears that the

number of categories influences the percentage of bias. Holding everything constant but increasing the number of response categories from 2 to 4 reduces the bias estimates. Namely, under the MS_M model, similar to the results of MS_S, to achieve smaller item parameter bias, it is better to have larger sample sizes or more response categories for items. When the number of categories is low, increasing the sample size becomes "a necessity" according to the simulation results. It is also important to note that all estimates of the bias were still larger than the 5% critical value.

Besides biases in parameter estimation, model-data fit indices were also examined under misspecified models (MS_S and MS_M). Specifically, the p-values of the chi-square test, CFI, RMSEA, and SRMR values were checked.

Figure 4. *Bias of the parameter estimates for MS_M.*

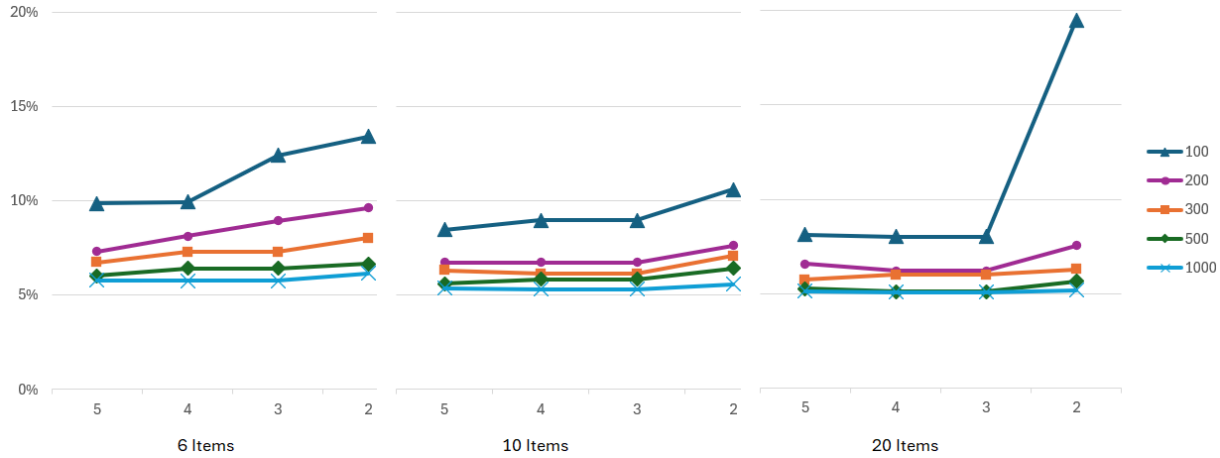


Table 1 shows the percentage of model fits for each CFI, RMSEA, and SRMR. The models reported in Table 1 are misspecified: The correlations between the factors were .5 and .8, respectively. Therefore, the values in the table were expected to be small. As the numbers increased, detecting misspecified models became less likely. For example, with 6 items, a sample size of 100, and a .5 correlation, the CFI value was .28. This means 28% of the data fit the model, even though the model was misspecified. The corresponding value for the .8 correlation (MS_M) was .98. Under the moderate misspecification with a .8 correlation, the percentage values increased. Most of the values were 1, indicating 100%. That is, 100% of the data showed fit even though it was a misspecified model. Based on the results, the values differed for the strong and moderate misspecification. It seems that when the misspecification was more evident (MS_S), as the number of categories decreased, the percentages of not rejecting the misspecified model increased based on the CFI and RMSEA values. However, as the sample size increases, up to 1000, this pattern disappears. For SRMR, the values remained almost the same regardless of size of response categories. However, under the small number of items, SRMR did not work properly, indicating biases larger than 5%, especially for the small sample sizes. Moreover, as the number of items increased, the values all became 0%. The pattern was different for the MS_M. The values were all much larger than the critical threshold for expected biases.

The results for the p-value of chi-square tests are given in Table 2. Similar to Table 1, the smaller the values, the better the misspecified model is detected. For example, with 6 items, sample size of 100, and a correlation of .5, the *p*-value was .10. This means 10% of the data fits the model according to the chi-square test, even though the model is misspecified. The corresponding value for a correlation of .8 (MS_M) was .89. When the corresponding values of MS_S and MS_M were compared, MS_M indicated larger values, in general. That is, a moderate level of misspecification is less likely to detect the problem. Based on the results, as the category decreased under the small item and sample size, detecting misspecification became less likely. As the number of items and/or sample size increased, the effect diminished.

Table 1. *Percent of model-data fit for misspecified models.*

Index	Number of Items	SS	MS_S				MS_M			
			C5	C4	C3	C2	C5	C4	C3	C2
CFI	6	100	.28	.33	.40	.46	.98	.99	.98	.96
		200	.14	.13	.22	.27	1.00	1.00	1.00	.99
		300	.06	.06	.11	.18	1.00	1.00	1.00	1.00
		500	.02	.01	.04	.07	1.00	1.00	1.00	1.00
		1000	.00	.00	.00	.01	1.00	1.00	1.00	1.00
	10	100	.26	.29	.41	.44	1.00	1.00	1.00	1.00
		200	.10	.12	.17	.22	1.00	1.00	1.00	1.00
		300	.03	.05	.08	.12	1.00	1.00	1.00	1.00
		500	.01	.01	.02	.03	1.00	1.00	1.00	1.00
		1000	.00	.00	.00	.00	1.00	1.00	1.00	1.00
	20	100	.24	.29	.37	.44	1.00	1.00	1.00	1.00
		200	.10	.10	.14	.20	1.00	1.00	1.00	1.00
		300	.03	.04	.04	.09	1.00	1.00	1.00	1.00
		500	.00	.01	.01	.02	1.00	1.00	1.00	1.00
		1000	.00	.00	.00	.00	1.00	1.00	1.00	1.00
RMSEA	6	100	.03	.05	.20	.31	.70	.76	.87	.87
		200	.00	.00	.03	.11	.67	.72	.85	.88
		300	.00	.00	.01	.03	.60	.66	.83	.91
		500	.00	.00	.00	.01	.48	.58	.84	.92
		1000	.00	.00	.00	.00	.31	.49	.86	.96
	10	100	.01	.03	.18	.29	.85	.90	.95	.96
		200	.00	.00	.02	.08	.81	.88	.97	.98
		300	.00	.00	.00	.02	.80	.87	.99	1.00
		500	.00	.00	.00	.00	.76	.85	.98	1.00
		1000	.00	.00	.00	.00	.68	.84	1.00	1.00
	20	100	.01	.01	.12	.26	.95	.98	1.00	1.00
		200	.00	.00	.00	.06	.96	.99	1.00	1.00
		300	.00	.00	.00	.01	95	.99	1.00	1.00
		500	.00	.00	.00	.00	.94	.99	1.00	1.00
		1000	.00	.00	.00	.00	.93	.99	1.00	1.00
SRMR	6	100	.07	.08	.07	08	.90	.87	.65	.51
		200	.07	.06	.06	.07	.99	.98	.94	.86
		300	.05	.04	.04	.05	1.00	1.00	.98	.95
		500	.02	.01	.04	.04	1.00	1.00	1.00	.99
		1000	.00	.00	.01	.01	1.00	1.00	1.00	1.00
	10	100	.00	.00	.00	.00	.84	.74	.32	.13
		200	.00	.00	.00	.00	.99	.98	.87	.69
		300	.00	.00	.00	.00	1.00	1.00	.99	.92
		500	.00	.00	.00	.00	1.00	1.00	1.00	1.00
		1000	.00	.00	.00	.00	1.00	1.00	1.00	1.00
	20	100	.00	.00	.00	.00	.69	.50	.03	.00
		200	.00	.00	.00	.00	.99	.98	.80	.37
		300	.00	.00	.00	.00	1.00	1.00	.99	.88
		500	.00	.00	.00	.00	1.00	1.00	1.00	1.00
		1000	.00	.00	.00	.00	1.00	1.00	1.00	1.00

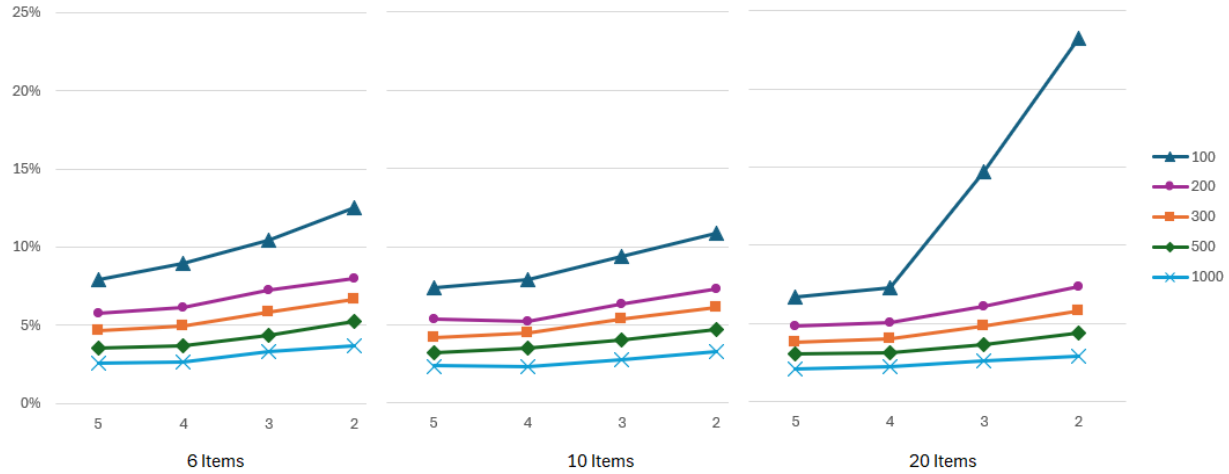
Table 2. *Percent of model-data fit based on the p-value for misspecified models.*

Number of Items	SS	MS_S				MS_M			
		C5	C4	C3	C2	C5	C4	C3	C2
6	100	.10	.19	.40	.55	.89	.92	.96	.96
	200	.00	.01	.04	.14	.72	.78	.88	.92
	300	.00	.00	.00	.02	.04	.09	.09	.48
	500	.00	.00	.00	.00	.00	.00	.00	.12
	1000	.00	.00	.00	.00	.00	.00	.00	.00
10	100	01	.04	.20	.34	.22	.30	.30	.88
	200	.00	.00	.01	.02	.00	.00	.00	.57
	300	.00	.00	.00	.00	.00	.00	.00	.26
	500	.00	.00	.00	.00	.00	.00	.00	.01
	1000	.00	.00	.00	.00	.00	.00	.00	.00
20	100	.00	.00	.04	.10	.02	.03	.03	.84
	200	.00	.00	.00	.00	.00	.00	.00	.41
	300	.00	.00	.00	.00	.00	.00	00	.05
	500	.00	.00	.00	.00	.00	.00	.00	.00
	1000	.00	.00	.00	.00	.00	.00	.00	.00

3.2. Results Based on Correctly Specified Models

Percentages of bias for the correctly specified model (CS) were given in Figure 5. Under the CS model, the data were generated with a correlation of 1 between factors (see Figure 1). As expected, as the sample size increases the percentage bias decreases, regardless of the test length or number of categories. Although the effect is limited, increasing the number of items contributes to a reduction in bias. For example, with a sample size was 200 and five response categories, biases were 5.8, 5.4, and 4.8 for 6, 10, and 20 item models respectively. It is worth noting that as the number of items reached 20, the bias dropped below the 5% critical value. Additionally, the number of response categories influenced bias estimates. Specifically, as the number of categories decreased, bias increased, regardless of the item and sample sizes.

Figure 5. *Bias of the parameter estimates for CS.*



Increasing sample sizes or the number of items eventually reduces the percent bias values below the critical value. Furthermore, increasing the number response options also reduces bias. Percentages of model data fit for the correctly specified model were reported in Table 3. The values in the table were expected to be small, indicating the percentages of non-fit for the model. For example, with 6 items and a sample size of 100, the RMSEA value was .02. This means that only 2% of the data does not fit the model. However, when viewed comprehensively, most

of the values in Table 3 were zero. Non-zero values were visible only for the sample size of 100. Based on the results, only SRMR values distinguish model-data fit. CFI and RMSEA values were almost all zero, while non-zero values were smaller than .05. Also, the *p*-value of the Chi-square test was all zero. Therefore, Chi-square, CFI, and RMSEA do not provide any feedback about model-data fit. However, SRMR indicates some non-zero values.

Table 3. *Percent of model-data fit for correctly specified models.*

Index	Number of Items	SS	C5	C4	C3	C2
CFI	6	100	0	0	0	0
		200	0	0	0	0
	10	100	0	0	0	0
		200	0	0	0	0
	20	100	0	0	0	0
		200	0	0	0	0
RMSEA	6	100	.02	.01	.01	.03
		200	0	0	0	0
	10	100	0	0	0	0
		200	0	0	0	0
	20	100	0	0	0	0
		200	0	0	0	0
SRMR	6	100	0	0	.06	.18
		200	0	0	0	0
	10	100	0	0	.16	.52
		200	0	0	0	0
	20	100	0	0	.43	.95
		200	0	0	0	0

Note: Since the values were zero (0) for sample sizes of 300, 500, and 1000, these cases were omitted from the table for clarity.

All these non-zero values corresponded to samples with sizes below 100, with the numbers of categories being three and two. That is if all three supplementary fit indices were examined together, under the small sample sizes and a smaller number of response categories, model data fit points to a problem. Therefore, if the sample size is small, increasing the number of response categories can solve model data fit issues. From another point of view, as long as an adequate sample size is given (larger than 200), the number of response categories does not have a direct impact on model-data fit. Furthermore, as the model complexity (*df*) increases, the effect on model-data fit increases as well. For instance, in a structure with six items and two response categories, model-data fit was not achieved 18% of the time, whereas, with 10 and 20 items, this rate rises to 52% and 95%, respectively.

4. DISCUSSION and CONCLUSION

The effects of the number of categories on factor analysis were examined in this study. For this purpose, the Monte Carlo simulation technique was used. Sample size, test length, response categories, and model specifications were used as the design factor of the simulation. Under a misspecified model, considering that the required sample size is expected to increase as the number of items grows, a sample of 100 for 20 items may be regarded as rather low. Furthermore, in such a small sample, reducing the number of response categories appears to

have significantly impacted the estimation of model parameters. Consequently, having fewer categories changes the results considerably until the sample size reaches a certain threshold. To reduce bias, it would be beneficial either to increase the test length or the number of categories, in cases where an adequate sample size is not feasible due to the study population. More broadly, regardless of sample size, increasing the number of categories can at the very least prevent an increase in bias under a misspecified model. On the other hand, when the number of categories is low, increasing either the sample size or the number of items, even slightly, can help reduce bias. Therefore, the number of categories is linked to both the sample and test length requirements.

For model-data fits, it is better to evaluate MS_S and MS_M separately. Moderate levels of misspecification were likely more challenging for the model to identify, as they more closely resembled the true structure. In other words, when the correlation is as high as .8, distinguishing a single-factor structure in the predictive model becomes more challenging. Thus, in assessing misspecification, it would be more appropriate to focus on the result of MS_S.

As with supplementary fit indices such as SRMR, the number of response categories also impacts chi-square values (Shi *et al*., 2021). Specifically, when sample size and test length are low, reducing the number of categories makes it harder to correctly identify a misspecified model (Maydeu-Olivares *et al*., 2017). In this context, in terms of chi-square values, increasing the number of categories can make a positive contribution to the identification of a misspecified model (Abulela & Khalaf, 2024), especially when sample size and test length are limited.

In sum, when estimating a model that is not correctly specified (such as MS_S), it becomes essential to increase either the number of items in the model or the sample size if the number of categories is low. Put differently, if a small number of response options is to be used when designing items, it is important to ensure a larger number of items or a substantial sample size. Therefore, in cases where item development or access to samples is challenging, increasing the number of categories becomes crucial to accurately estimating model-data fit in misspecified models.

Similar to misspecified models, the results showed that, for the correctly specified model, increasing the sample size (MacCallum *et al*., 1999) or the number of items (Simms *et al*., 2019) reduces bias values. Furthermore, results indicate that the number of response categories used in items also helps lower estimated bias values (Shi *et al*., 2021). Consequently, in situations where sample access is limited or the number of items is restricted, increasing the number of response options in categorically defined items serves as a compensating factor. Conversely, if the number of response options is reduced, achieving consistent results would require either more items or a larger sample size. In a nutshell, when the sample size and test length are limited, choosing more response options can help lower the bias estimates.

In a correctly specified model, having a sample size above 200 can help prevent potential issues related to model-data fit. However, when the sample is insufficient or challenging to obtain, using items with only two or three response categories can lead to model-data misfit, at least as indicated by the SRMR value. Considering that supplementary fit indices are generally evaluated together, a low number of response categories may falsely suggest problems with model-data fit.

In conclusion, findings show that in misspecified models, low sample sizes and fewer response categories increase bias and lower the accuracies of parameter estimation (Lozano *et al*., 2008, Shi *et al*., 2021). To mitigate these effects, larger samples or more response categories are recommended. Particularly in models with small sample sizes and/or item counts, the number of response categories must be increased to reduce bias. Correctly specified models also benefit from additional categories, reducing bias and supporting model-data fit. A sample size of over 200 is ideal for avoiding model misfit, as low category counts with limited samples may inaccurately indicate poor fit. Finally, the study demonstrates that sample size requirements are

influenced not only by factors such as study design and estimator type (Kyriazos, 2018) but also by the number of response categories in the items. Specifically, in studies employing categorical data, using a larger number of response categories (e.g., 5) rather than fewer (e.g., 2 or 3) positively affects the adequacy of the sample size. This suggests that employing a greater number of response options can help meet sample size requirements more effectively.

Future research could examine scenarios involving studies with more than five categories, as this study is limited to a maximum of five. Furthermore, it would be beneficial to analyze how outcomes might vary in the context of more complex models. Finally, evaluating the use of a different estimation method could yield valuable insights.

Declaration of Conflicting Interests and Ethics

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

Orcid

Fatih Orçan https://orcid.org/0000-0003-1727-0456

REFERENCES

Abulela, M.A.A., & Khalaf, M.A. (2024). Does the number of response categories impact validity evidence in self-report measures? A scoping review. *Sage Open, 14*(1), 1-16. https://doi.org/10.1177/21582440241230363

Abdelsamea, M. (2020). The effect of the number of response categories on the assumptions and outputs of item exploratory and confirmatory factor analyses of measurement instruments in psychological research. *Journal of Education Sohag UNV, 76*, 1153-1222. https://doi.org/10.21608/edusohag.2020.103373

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. https://doi.org/10.1207/s15324818ame0902_4

Flora, D.B., & Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491. https://doi.org/10.1037/1082-989X.9.4.466

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kline, R.B. (2011). *Principles and Practice of Structural Equation Modeling*. Guilford Press.

Komorita, S.S., & Graham, W.K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 25*(4), 987-995. https://doi.org/10.1177/001316446502500404

Kılıç, A.F. (2022). The effect of categories and distribution of variables on correlation coefficients. *Ege Eğitim Dergisi, 23*(1), 50-80. https://doi.org/10.12984/egeefd.890104

Kyriazos, T.A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology, 9*(8), 2207-2230. https://doi.org/10.4236/psych.2018.98126

Lozano, L., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology 4*(2). 73-79. https://doi.org/10.1027/1614-2241.4.2.73

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Matell, M.S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? I. Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674. https://doi.org/10.1177/001316447103100307

Maydeu-Olivares, A., Fairchild, A.J., & Hall, A.G. (2017). Goodness of fit in item factor analysis: Effect of the number of response alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 495-505. https://doi.org/10.1080/10705511.2017.1289816

Orçan, F. (2021). MonteCarloSEM: An R package to simulate data for SEM. *International Journal of Assessment Tools in Education*, *8*(3), 704-713. https://doi.org/10.21449/ijate.804203

Orçan, F., & Yanyun, Y. (2016). A note on the use of item parceling in structural equation modeling with missing data. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(1), 59-72. https://doi.org/10.21031/epod.88204

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Shi, D., Siceloff, E.R., Castellanos, R.E., Bridges, R.M., Jiang, Z., Flory, K., & Benson, K. (2021). Revisiting the effect of varying the number of response alternatives in clinical assessment: Evidence from measuring ADHD symptoms. *Assessment*, *28*(5), 1287-1300. https://doi.org/10.1177/1073191120952885

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*(4), 557-566. http://dx.doi.org/10.1037/pas0000648

Yoon, G. (2024) No one optimal way to measure people's attitudes? Preferred length of scales in advertising research. *Journal of Current Issues & Research in Advertising*, *45*(1), 43-70, https://doi.org/10.1080/10641734.2023.2246049

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement*, *72*(4), 533-546. https://doi.org/10.1177/0013164411431162