

# Examination of ChatGPT's Performance as a Data Analysis Tool



Duygu Koçak<sup>1</sup> 

## Abstract

This study examines the performance of ChatGPT, developed by OpenAI and widely used as an AI-based conversational tool, as a data analysis tool through exploratory factor analysis (EFA). To this end, simulated data were generated under various data conditions, including normal distribution, response category, sample size, test length, factor loading, and measurement models. The generated data were analyzed using ChatGPT-4o twice with a 1-week interval under the same prompt, and the results were compared with those obtained using R code. In data analysis, the Kaiser–Meyer–Olkin (KMO) value, total variance explained, and the number of factors estimated using the empirical Kaiser criterion, Hull method, and Kaiser–Guttman criterion, as well as factor loadings, were calculated. The findings obtained from ChatGPT at two different times were found to be consistent with those obtained using R. Overall, ChatGPT demonstrated good performance for steps that require only computational decisions without involving researcher judgment or theoretical evaluation (such as KMO, total variance explained, and factor loadings). However, for multidimensional structures, although the estimated number of factors was consistent across analyses, biases were observed, suggesting that researchers should exercise caution in such decisions.

## Keywords

exploratory factor analysis, artificial intelligence, ChatGPT, data analysis, accuracy estimation percentage, relative bias

---

<sup>1</sup>Alanya Alaaddin Keykubat University, Alanya/Antalya, Turkey

## Corresponding Author:

Duygu Koçak, Department of Measurement and Assessment, Alanya Alaaddin Keykubat University, Antalya 07425, Turkey.

Email: [duygu.kocak@alanya.edu.tr](mailto:duygu.kocak@alanya.edu.tr)

## **Introduction**

With the advancement of generative artificial intelligence (AI), AI applications are being widely used across various fields, and among the most popular of these is ChatGPT. ChatGPT is a natural language processing tool powered by AI technology, allowing users to interact with a chatbot. Over the past few years, this language model has frequently been used to perform various tasks and answer questions (Baidoo-Anu & Owusu Ansah, 2023; Mollick & Mollick, 2022; Rudolph & Tan, 2023). ChatGPT was launched by OpenAI on November 30, 2022, and since then, the click-through rate to OpenAI has increased by approximately 3,500%, reaching 672 million visitors from 18.3 million (Cribben & Zeinali, 2023). Some of these users are curious about the capabilities of AI, while others actively access the application.

ChatGPT is built on a large language model (LLM) architecture known as the Generative Pre-trained Transformer (GPT; Dehouche, 2021; Liang et al., 2023). GPT-3 and its more advanced versions, GPT-3.5 and GPT-4, are autoregressive language models that leverage deep learning to generate human-like texts. These models have been trained on a vast amount of text data sourced from books, articles, blogs, and various websites (Kasneji et al., 2023; Mhlanga, 2023; Tlili et al., 2023). The language model that forms the basis of ChatGPT is designed with a supervised and reinforcement learning approach, which OpenAI refers to as “Reinforcement Learning with Human Feedback” (RLHF). According to OpenAI, the key factor that distinguishes ChatGPT from other AI applications and makes it powerful is the use of this RLHF method (Singh et al., 2022).

On March 14, 2023, OpenAI introduced GPT-4, available through ChatGPT Plus, which surpasses other models in many ways. GPT-4 is more reliable, creative, and capable of handling complex instructions than GPT-3.5. With 170 trillion parameters compared with GPT-3’s 175 billion, GPT-4 is significantly larger and more powerful. According to OpenAI, users will notice the difference between GPT-3.5 and GPT-4 more clearly when they attempt to perform complex tasks. The most advanced GPT-based model, Auto-GPT, was published on GitHub by Significant Gravitas on March 30, 2023. This model, an open-source Python application powered by GPT-4, is capable of completing tasks with minimal human intervention. The main difference between Auto-GPT and ChatGPT is that Auto-GPT can operate independently without human assistance, whereas ChatGPT requires human commands to function effectively.

ChatGPT is used in various academic and non-academic fields to analyze inputs and take actions. For instance, it can analyze data from sensors or other monitoring devices to provide real-time information about a patient’s health status, serving as a tool for remote patient monitoring (Iftikhar, 2023). Alshurafat (2023) has examined how ChatGPT can transform the work systems of employees in accounting, improving their efficiency and productivity. Moreover, ChatGPT has demonstrated the ability to write short stories based on the information provided by users (McGee, 2023). In addition to fields like health care, finance, and education, commercial enterprises also benefit from this technology. For example, companies use ChatGPT for purposes

such as generating marketing content, developing different marketing strategies, writing computer code, and providing post-sales customer service.

AI has become so advanced that some researchers argue that ChatGPT could succeed in undergraduate studies. Terwiesch (2023) initiated a discussion by focusing on whether ChatGPT could earn an MBA degree from Wharton. Similarly, Wood et al. (2023) examined how well ChatGPT answers accounting assessment questions. However, ethical issues such as plagiarism and the accuracy of the information provided by ChatGPT are also subjects of debate (Gillani et al., 2023; Nemorin et al., 2023).

Since ChatGPT's advanced versions can perform a wide range of tasks, it is used by users for various purposes. Especially with the release of GPT-4, more people began to use this technology because it offers users the ability to carry out more reliable, creative, and complex instructions. Auto-GPT, the most advanced open-source Python application based on GPT, was published by Significant Gravititas and is supported by GPT-4. This means that GPT-4 can analyze data when provided in an appropriate format. In other words, the GPT-4 version of ChatGPT is capable of performing data analysis. This development enables researchers to use AI for data analysis in scientific studies.

ChatGPT performs statistical analyses by integrating external tools like Python. Based on the data and instructions provided to it, ChatGPT generates Python code (Frantz, 2024) and then executes this code to conduct statistical analyses. For instance, it can perform analyses like *t*-tests and factor analyses using Python libraries such as pandas, scipy, and statsmodels. These processes automate analyses that a data scientist would typically perform manually. Through its Advanced Data Analysis feature, ChatGPT can read, clean, analyze, and visualize datasets. It can create graphs using Python tools like matplotlib or seaborn. However, these analytical capabilities face limitations when data size or complexity increases. Working with large datasets can be challenging, and the results may sometimes be incomplete or inaccurate; certain prompts may not be understood, particularly if they involve technical terms and explanations (Ambati et al., 2024; Latendresse et al., 2024).

Data analysis is a crucial step in scientific research. Collecting data appropriate to the research question, choosing the correct statistical method, and obtaining accurate findings are vital for the reliability, internal validity, and external validity of the study. The challenges associated with data analysis have been largely overcome with the advent of computers in scientific research and the emergence of statistical software packages (Field, 2009). Statistical analyses that were once difficult or time-consuming can now be easily conducted using computers and software (Kalaycı, 2010; Tan, 2016). Although a researcher does not have to be a statistician, they need knowledge of statistical techniques to analyze the data correctly (Erkuş, 2011). In addition, the researcher should be able to select the appropriate statistical analysis (Creswell, 2012), organize the research data accordingly, and apply the commands accurately in the program (Pallant, 2005). Thus, a researcher must know both which statistical analysis to perform for the research question and how to interpret the

analysis results and use the necessary software. Erkuş (2011) emphasized that simply having the technical skill to use software packages is insufficient for data analysis; statistical analyses without understanding variable structure and procedures may yield erroneous results. Therefore, researchers need both statistical knowledge and software proficiency. This situation poses a challenge for researchers who require statistical analyses as a tool in their research, as it demands both statistical and programming knowledge.

ChatGPT, in this context, can perform analyses and report findings if the researcher specifies the type of analysis needed and provides the data in an appropriate format (a guide detailing the steps of analysis using ChatGPT is presented in the Appendix). This offers significant convenience for researchers; however, how can we be certain of the accuracy of ChatGPT's results? The best way to address this question is to examine ChatGPT's performance in statistical analyses. Given that researchers in fields such as health care, education, psychology, and business might wish to use ChatGPT for data analysis, it would be appropriate to assess the application's performance using a universally applicable statistical technique.

It is evident that ChatGPT is capable of performing various analyses. At this point, whether the results produced by ChatGPT are biased or accurate is a critical question for users and deserves examination. Therefore, this study aims to evaluate ChatGPT's performance as a data analysis tool. Specifically, ChatGPT's effectiveness is assessed through exploratory factor analysis (EFA), a statistical technique frequently used in social sciences.

When considering the requirements of EFA, it can be stated that some steps involve technical comparisons for decision-making (e.g., determining if the Kaiser–Meyer–Olkin [KMO] value is adequate for factorization, and defining the number of factors), while others require decisions that take theoretical frameworks into account. The extent to which an AI tool can accurately perform these steps is a critical question for both literature and researchers. Thus, this study aims to examine ChatGPT's performance as a data analysis tool by analyzing factor analysis parameters.

### ***Research Model***

In this study, the performance of the artificial intelligence tool ChatGPT in conducting factor analysis was examined using artificially generated data. As the study involves comparing the performance of a program and software using simulated data, it aligns with the characteristics of a simulation study (Dooley, 2002).

### ***Simulation Conditions***

In this study, which examines ChatGPT's data analysis performance through factor analysis, variables such as the number of items, sample size, factor loading, and measurement model were manipulated.

The required sample size for factor analysis should be determined by considering model-related parameters, such as the number of factors, the number of items, and the relationship between factors in the analyzed model. Muthen and Muthen (2002) stated that for confirmatory factor analysis with two factors, a factor interrelation of 0.25, a factor loading of 0.81, and an error variance of 0.36, the necessary sample size is 150, and as the model complexity increases, the required sample size also increases. For determining the necessary sample size for EFA in this study, the R packages MBESS (Kelley et al., 2018) and “pwr” (Champely et al., 2018) were used. Statistical power was set to 0.80, and the necessary sample size for each condition was calculated based on the manipulated variables: item count, number of factors, factor interrelationship, and factor loading. The code used for these calculations is presented as Online Supplementary Materials (OSF. <https://osf.io/xk59t>). The minimum sample size required was 201 for a one-dimensional condition with 10 items. The maximum sample size needed was 225 for a two-dimensional condition with 10 items and a factor interrelationship of 0.55. Considering these results, a sample size of 225 was chosen for data generation in this study.

Recommendations regarding the necessary number of items for forming a factor (Ryan et al., 2018; Siow et al., 2017; Tang et al., 2022) were taken into account, and the test length was set at 10 items in the generated data. Scoring with five categories was used, as it is suggested to better reveal psychometric properties (Adelson & McCoach, 2010; Mellor & Moore, 2014); thus, response categories were fixed at five. Based on the assumption that psychological traits are normally distributed in the population, which is one of the basic assumptions in psychometrics (Crocker & Algina, 2008; Ho & Yu, 2015), data distributions were set to normal.

The study examined related multidimensional structures, unrelated multidimensional structures, and unidimensional structures. While most psychological constructs are multidimensional, unidimensional structures are also encountered (Nunnally & Bernstein, 1994). Therefore, both unidimensional and two-dimensional structures were considered in this study. The related multidimensional structure was modeled to have a moderate interrelation of 0.50. In the literature, many studies fix item factor loadings to a single value (Beauducel & Herzberg, 2006; Li, 2016). It is suggested that for an item to be considered significant in the factor it loads on, its factor loading should be at least 0.30. Based on this, values in Table 1 were established to set items' communalities, extractions, and factor loadings. Factor loadings were fixed between 0.50 and 0.70 for the unidimensional model, and between 0.60 and 0.85 for the related and unrelated two-dimensional models.

In this study, the simulation conditions included sample size (1), test length (1), number of categories for item scores (1), distribution of item scores (1), and measurement model (3), resulting in  $1 \times 1 \times 1 \times 1 \times 3 = 3$  conditions. Each condition was analyzed with 100 repetitions, yielding analyses on 300 datasets. According to Harwell et al. (1996), at least 25 replications are recommended for Monte Carlo simulation studies to ensure generalizable results. In this study, data were generated with 100 repetitions to enhance the robustness and reliability of the findings.

**Table 1.** Simulation Conditions.

Factor		Condition		
Sample size		225		
Test Length		10		
Number of Categories for Item Scores		5		
Distribution of Item Scores		Normal		
Measurement Model	Unidimensional Model	Two-Dimensional Model		
		<ul style="list-style-type: none"> <li>• Related (<math>\varphi = 0.50</math>)</li> <li>• Unrelated (<math>\varphi=0.00</math>)</li> </ul>		
		Unidimensional model	Two-dimensional model	
			$\varphi=0.50$	$\varphi=0.00$
Factor Loading (FL)	Items	FL	FL	FL
	Item1	0.56	0.71	0.80
	Item2	0.64	0.72	0.79
	Item3	0.71	0.81	0.73
	Item4	0.56	0.62	0.66
	Item5	0.59	0.67	0.60
	Item6	0.67	0.77	0.73
	Item7	0.62	0.75	0.67
	Item8	0.64	0.78	0.67
	Item9	0.64	0.79	0.67
	Item10	0.58	0.80	0.64

### Data Generation

For each condition in the study, continuous datasets were generated using the R packages *lavaan* (Rosseel, 2012) and *MASS* (Ripley et al., 2013). The code used for these calculations is presented as Online Supplementary Materials (OSF. <https://osf.io/xk59t>).

### Data Analysis

Following data generation, EFA was conducted separately on the generated datasets using both R software and ChatGPT-4. For both R and ChatGPT-4 analyses, the maximum likelihood estimation method was employed as the basis for the EFA.

In R, a loop was incorporated into the code to automate the analysis of each generated dataset, allowing for efficient repetition of factor analyses across all datasets. For analyses conducted with ChatGPT, each dataset was saved separately and analyzed individually by inputting each dataset one at a time into ChatGPT. Since ChatGPT cannot directly retrieve findings from R or its packages, this process was manually facilitated by researchers.

*Analysis Process with R.* For the analyses conducted with R (R Development Core Team, 2011), the following packages were used: *psych* (Revelle & Revelle, 2015), *GPArotation* (Bernaards et al., 2015), *nFactors* (Raiche et al., 2020), *paran* (Dinno & Dinno, 2018), *lavaan* (Rosseel et al., 2017), *semPlot* (Epskamp, 2017), and *EFA.MRFA* (Navarro-Gonzalez et al., 2020). The analyses included the following steps:

- Suitability for EFA: Assessed using the KMO coefficient to determine if the data were suitable for EFA.
- Number of Factors: Determined using the empirical Kaiser criterion, Hull method, and Kaiser–Guttman criterion.
- Total Variance Explained: Calculated to understand the proportion of variance accounted for by the extracted factors.
- Factor Loadings: Calculated for each item to assess its loading strength on the identified factors.
- In addition, rotations were performed using both the varimax (orthogonal) and promax (oblique) methods to allow for a comparison of factor loadings post-rotation.

The code used for these calculations is presented as Online Supplementary Materials (OSF. <https://osf.io/xk59t>). Since seven different R packages were utilized, “R” is used as a general reference to the implementation in the reporting rather than listing each package individually. However, it should be noted that the results were obtained using the specific code outlined in the Online Supplementary Materials.

*Analysis Process With ChatGPT-4.* ChatGPT performs statistical analyses by integrating external tools like Python. It generates Python code based on the provided data and instructions (Frantz, 2024) and then executes this code to carry out statistical analyses. With its Advanced Data Analysis (ADA) feature, ChatGPT can read, clean, analyze, and visualize datasets. Therefore, when provided with datasets in a compatible format and clear, accurate prompts, ChatGPT can integrate Python libraries—such as pandas, scipy, and statsmodels—to conduct the requested analyses.

Given that AI applications may produce varying responses to identical prompts at different times (Tassoti, 2024; Theophilou et al., 2023), the same analysis was conducted on the same data with the same prompt, once a week apart, to examine any differences in the results. The findings from the first analysis are reported as ChatGPT\* in the results section, while the second analysis, conducted a week later, is denoted as ChatGPT\*\*.

The datasets generated in R were provided to ChatGPT-4 as CSV files, with instructions to perform factor analysis and calculate specific parameters. Appropriate prompts were crafted to ensure that ChatGPT interpreted and returned the requested calculations accurately. The requested parameters included:

- KMO Coefficient: To assess the suitability of the data for factor analysis.
- Number of Factors: Estimated using the empirical Kaiser criterion, Hull method, and Kaiser–Guttman criterion.
- Total Variance Explained: To evaluate the cumulative variance accounted for by the extracted factors.
- Factor Loadings: Calculated for each item.

In addition, rotations were performed using both varimax (orthogonal) and promax (oblique) methods to compare factor loadings post-rotation. This comprehensive setup allowed for a robust comparison of ChatGPT's performance in factor analysis and its consistency with conventional statistical software. The generated data were analyzed individually in ChatGPT. Subsequently, the analysis findings were recorded and consolidated.

### *Evaluation of Analysis Outputs*

The KMO values, total variance explained, and item factor loadings after rotation obtained via R and ChatGPT were descriptively compared. The estimated number of factors and average factor loadings, determined using the empirical Kaiser criterion, Hull method, and Kaiser–Guttman criterion, were compared to the predefined simulation values using relative bias ratio (RBR) and accuracy estimation percentage (AEP).

The distance of the estimated values from the true values is essential for the validity, power of the simulation study (Collins et al., 2001; Flora & Curran, 2004), and the interpretation of the findings. Thus, the accuracy of the factor number and average factor loading estimates provided by R and ChatGPT-4 was examined in comparison with the criteria set during data generation.

- AEP: This metric expresses, as a percentage, how often the estimated factor number and factor loading align with the true factor number and average factor loading in the simulations. Higher AEP values, approaching 100, indicate more accurate estimates. A tolerance of up to 10% error is considered acceptable; values exceeding this indicate a deviation from the simulated conditions, suggesting inaccurate estimates (Collins et al., 2001).
- RBR: RBR quantifies the deviation of estimated averages from the simulation values, indicating potential bias in the estimates. The RBR, calculated using the formula,  $(\hat{\theta} - \theta_{\text{real}}) / \theta_{\text{real}}$ , where  $\hat{\theta}$  represents the mean estimate obtained from 100 repetitions (in this study, it refers to the average number of factors and item factor loading), and  $\theta_{\text{real}}$  represents the parameter specified in the simulation (in this study, factor numbers of 1 and 2, and factor loading values provided in Table 1). The presence of either a negative (underestimation) or a positive (overestimation) bias in factor numbers or factor loadings is

undesirable. The literature suggests an acceptable bias range of  $|RBR| \leq 0.10$  (moderate bias) (Flora & Curran, 2004; Moshagen & Musch, 2014).

The code segments for calculating the AEP and RBR for the R-derived results are presented as Online Supplementary Materials (OSF. <https://osf.io/xk59t>). To calculate the AEP and RBR for ChatGPT's analysis findings, code was written in R using the *psych* (Revelle & Revelle, 2015), *GPArotation* (Bernaards et al., 2015), and *nFactors* (Raiche et al., 2020) packages.

## Results

In the study, the KMO values and total variance explained for the 10-item condition obtained from ChatGPT and R are presented in Table 2.

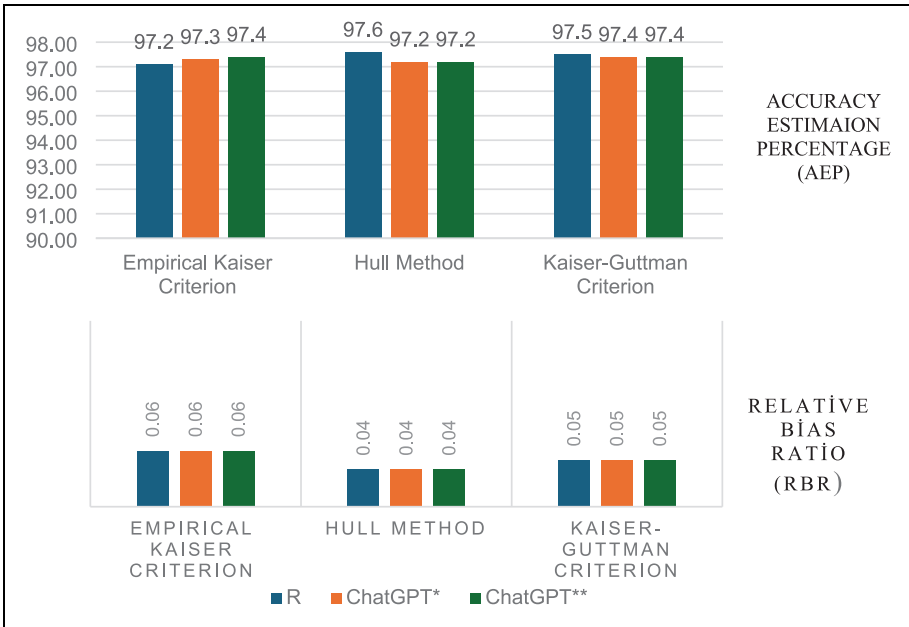
Examining Table 2 reveals that the KMO values and total variance explained are identical across analyses. This finding indicates that the EFA parameters provided by ChatGPT are consistent with those obtained using R. In addition, the results of the analyses conducted by ChatGPT, using the same prompt 1 week apart, are also consistent with each other.

In the study, AEP and RBRs were calculated for the number of factors determined by the empirical Kaiser criterion, Hull method, and Kaiser–Guttman criterion across single-factor, two-factor uncorrelated, and two-factor correlated structures. The AEP can be considered a measure of power for the study. Among the 100 replications conducted for each condition, the number of correctly and incorrectly estimated factor counts was recorded. As AEP values approach 100, it indicates that the method provides a more accurate estimation of factor numbers. For AEP values, an acceptable error margin of 10% was considered, meaning that if a method correctly estimates at

**Table 2.** KMO and Total Variance Explained for 10-Item Condition.

Model	n		KMO	Total variance explained
Single-Factor Model	225	R	0.915	39.90
		ChatGPT <sup>a</sup>	0.915	39.90
		ChatGPT <sup>b</sup>	0.915	39.90
Two-Dimensional Model (Correlated Factors)	225	R	0.923	68.13
		ChatGPT <sup>a</sup>	0.923	68.13
		ChatGPT <sup>b</sup>	0.923	68.13
Two-Dimensional Model (Uncorrelated Factors)	225	R	0.901	54.66
		ChatGPT <sup>a</sup>	0.901	54.66
		ChatGPT <sup>b</sup>	0.901	54.66

<sup>a</sup>The first analysis conducted by ChatGPT. <sup>b</sup> The analysis conducted 1 week later by ChatGPT using the same prompt.



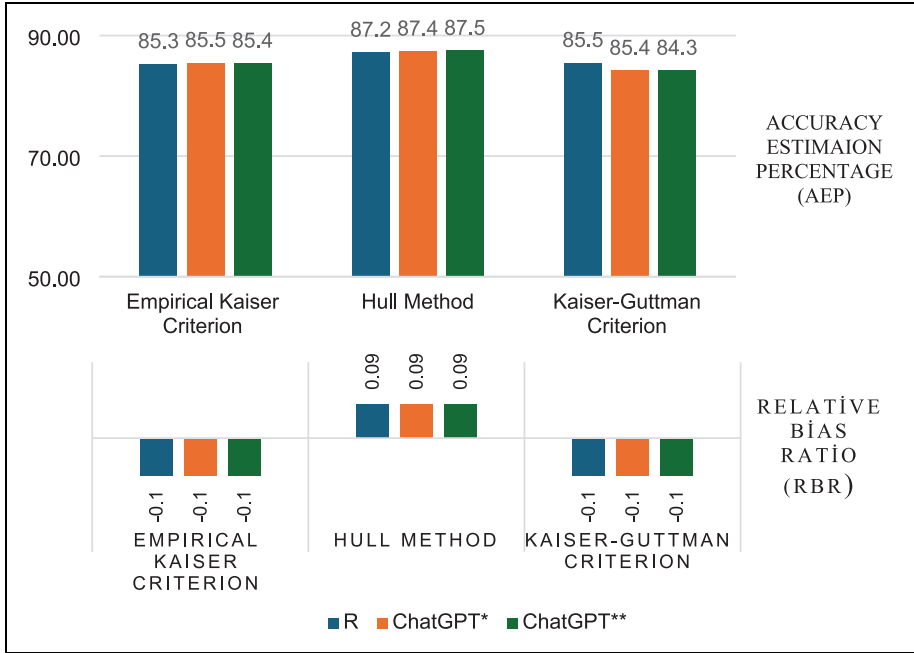
**Figure 1.** Accuracy Estimation Percentage (AEP) and Relative Bias Ratio (RBR) for Factor Count in Single-Factor Models.

least 90 out of 100 replications for a given condition, it is deemed acceptable (Collins et al., 2001).

Accordingly, the number of factors in the generated data was first estimated using the selected methods. Subsequently, the AEPs and RBRs for the estimated factor counts were calculated for single-factor, two-factor uncorrelated, and two-factor correlated models using both R and ChatGPT.

Figure 8 displays the AEPs and RBRs obtained for the single-factor model. For this model, the factor count is 1. The empirical Kaiser criterion, Hull method, and Kaiser–Guttman criterion were applied to all data generation scenarios involving a single-factor structure, with the alignment of factor counts estimated by R and ChatGPT with the actual factor count taken into account in the calculation.

Examining Figure 1 reveals that the AEP and RBR for factor count calculations in single-factor data are very similar across analyses conducted by R and ChatGPT, even when ChatGPT’s analyses were repeated 1-week apart. This suggests that ChatGPT’s factor analysis results are consistent over time and align closely with the findings obtained from R.



**Figure 2.** Two-Dimensional Uncorrelated Structure.

In addition, given that the AEP exceeds 90% and the RBR is below 0.10, it can be concluded that the factor count estimation in the single-factor model shows no bias and achieves high accuracy for both ChatGPT and R.

In Figure 2, the AEP and RBR for factor count estimation were calculated for two-dimensional uncorrelated structures. The findings for the estimated factor count are consistent between R and ChatGPT, indicating that analyses conducted by ChatGPT on the same data with the same prompt at different times align closely with the results from R. Thus, it can be concluded that both methods yield the same outcome.

The AEP is above 90%, and the RBRs are below 0.1, suggesting that the factor determination methods used provide unbiased findings for two-dimensional uncorrelated models.

Examining Figure 3 reveals that the RBR and AEP for factor count in two-dimensional correlated structures are stable and consistent across findings from R and ChatGPT conducted at different times.

Figures 1–3 collectively demonstrate that the factor counts obtained from R and from ChatGPT analyses performed 1 week apart using the same prompt are similar across all conditions. Accordingly, the AEP and RBR values are either equal or very close, indicating that ChatGPT provides results consistent with R.

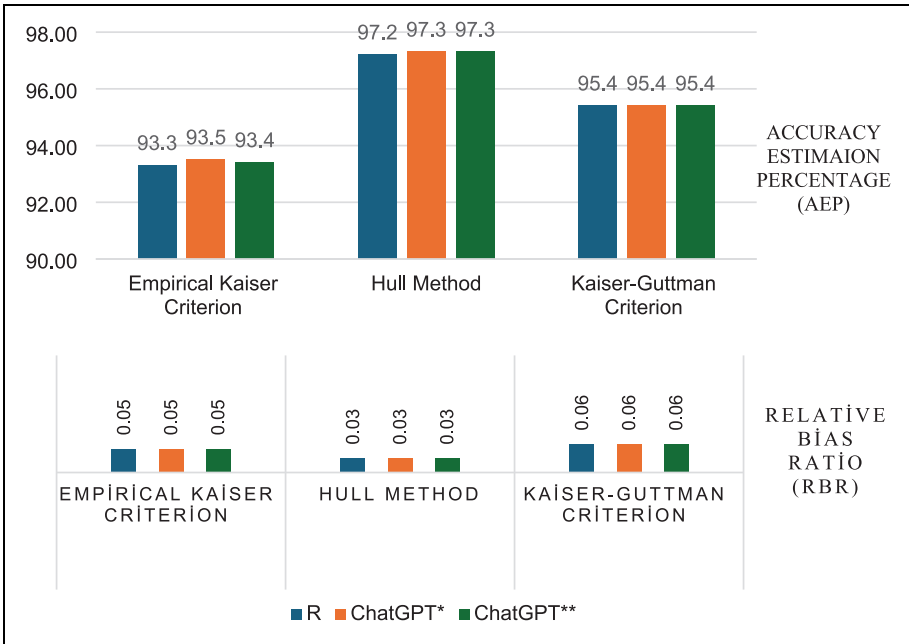


Figure 3. Two-Dimensional Correlated Structures.

Table 3. Factor Loadings in Single-Factor Model.

Item	Factor loadings		
	R	ChatGPT*	ChatGPT**
Item1	0.56	0.56	0.56
Item2	0.64	0.64	0.64
Item3	0.71	0.71	0.71
Item4	0.56	0.56	0.56
Item5	0.59	0.59	0.59
Item6	0.67	0.67	0.67
Item7	0.62	0.62	0.62
Item8	0.64	0.64	0.64
Item9	0.64	0.64	0.64
Item10	0.58	0.58	0.58

The estimates for item factor loadings in the single-factor model obtained from R and ChatGPT, with a 1-week interval between ChatGPT analyses, are presented in Table 3.

**Table 4.** Factor Loadings and Post-Rotation Factor Loadings in Two-Dimensional Uncorrelated Model.

Item	Initial factor solution						After rotation					
	R		ChatGPT**		ChatGPT**		Promax		Promax		Varimax	
	Dimensions		Dimensions		Dimensions		R		R		R	
	1	2	1	2	1	2	ChatGPT*	ChatGPT**	ChatGPT*	ChatGPT**	ChatGPT*	ChatGPT**
Item1	0.80	0.23	0.80	0.23	0.80	0.23	0.84	0.84	0.84	0.84	0.84	0.84
Item2	0.79	0.28	0.79	0.28	0.79	0.28	0.79	0.79	0.79	0.79	0.79	0.79
Item3	0.73	0.30	0.73	0.30	0.73	0.30	0.85	0.85	0.85	0.85	0.85	0.85
Item4	0.66	0.07	0.66	0.07	0.66	0.07	0.66	0.66	0.66	0.66	0.66	0.66
Item5	0.60	0.13	0.60	0.13	0.60	0.13	0.61	0.61	0.61	0.61	0.61	0.61
Item6	-0.17	0.72	-0.17	0.73	-0.17	0.72	0.75	0.75	0.75	0.75	0.75	0.75
Item7	-0.19	0.67	-0.19	0.67	-0.19	0.67	0.73	0.73	0.73	0.73	0.73	0.73
Item8	-0.32	0.67	-0.32	0.67	-0.32	0.67	0.66	0.66	0.66	0.66	0.66	0.66
Item9	-0.30	0.67	-0.30	0.67	-0.30	0.67	0.70	0.70	0.70	0.70	0.70	0.70
Item10	-0.14	0.64	-0.14	0.64	-0.14	0.64	0.73	0.73	0.73	0.73	0.73	0.73

Examining Table 3 reveals that R and ChatGPT provide identical values for item factor loadings. This consistency suggests that the estimates provided by ChatGPT, even when conducted 1 week apart, are aligned with each other and with the estimates obtained from R. Furthermore, these estimates are consistent with the factor loading values specified in the simulation conditions presented in Table 1. No rotation was performed for the single-factor model.

Examining Table 4 indicates that the factor analysis results from R and ChatGPT are identical and consistent with each other. Both Promax and Varimax rotations yield compatible results across R and ChatGPT, further demonstrating ChatGPT's similar performance to R.

As seen in Table 5, similar to the unidimensional model and two-dimensional uncorrelated model, R and ChatGPT yield comparable results both before and after rotation.

Table 6 presents the communalities for each item obtained through EFA conducted by R and ChatGPT across all three models. It is evident that R and ChatGPT produced consistent and identical results for each model.

The AEP and RBR were used to evaluate whether the factor loadings estimated through data analysis accurately reflected the conditions set in the simulation. These metrics provided insight into the degree to which the estimated factor loadings aligned with the predefined simulation parameters, indicating the reliability and precision of the estimates produced by R and ChatGPT.

The AEP and RBR calculated for factor loadings are presented in Figure 4. clicking on the paperclip icon allows you. The average factor loadings obtained from EFA conducted by ChatGPT (at 1-week intervals with identical prompts) and R are consistent with each other. The very low RBR indicates that neither R nor ChatGPT produced biased estimates for factor loadings. Similarly, the near-100% AEP confirms the accuracy of the average factor loading estimates provided by both R and ChatGPT, demonstrating the reliability of both methods in estimating factor loadings accurately.

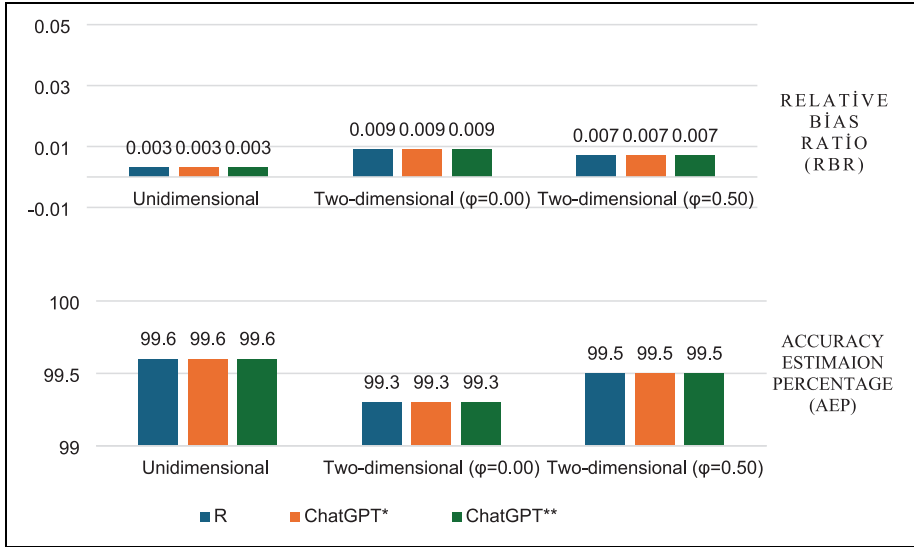
It was found that the KMO values, explained variance ratios, factor loadings, and communalities obtained from the analyses conducted by ChatGPT (at 1-week intervals) and R are consistent with each other. This indicates a high level of alignment between ChatGPT's estimates and those produced by R, confirming the reliability and accuracy of ChatGPT in performing EFA.

**Table 5.** Factor Loadings and Post-Rotation Factor Loadings in Two-Dimensional Correlated Model.

Item	Initial factor solution						After rotation					
	R		ChatGPT**		ChatGPT**		Promax		ChatGPT**		Varimax	
	Dimensions		Dimensions		Dimensions		R		ChatGPT**		R	
	1	2	1	2	1	2	1	2	ChatGPT*	ChatGPT**	ChatGPT*	ChatGPT**
Item1	0.71	0.43	0.71	0.43	0.71	0.43	0.89	0.89	0.89	0.89	0.81	0.81
Item2	0.72	0.18	0.72	0.18	0.72	0.18	0.62	0.62	0.62	0.62	0.64	0.64
Item3	0.81	0.32	0.81	0.32	0.81	0.32	0.82	0.82	0.82	0.82	0.80	0.80
Item4	0.62	0.38	0.62	0.38	0.62	0.38	0.78	0.78	0.78	0.78	0.70	0.70
Item5	0.67	-0.53	0.67	-0.53	0.67	-0.53	0.97	0.97	0.97	0.97	0.85	0.85
Item6	0.77	-0.44	0.77	-0.44	0.77	-0.44	0.93	0.93	0.93	0.93	0.86	0.86
Item7	0.75	-0.31	0.75	-0.31	0.75	-0.31	0.76	0.76	0.76	0.76	0.74	0.74
Item8	0.78	-0.18	0.78	-0.18	0.78	-0.18	0.63	0.63	0.63	0.63	0.67	0.67
Item9	0.79	-0.18	0.79	-0.18	0.79	-0.18	0.64	0.64	0.64	0.64	0.68	0.68
Item10	0.80	0.36	0.80	0.36	0.80	0.36	0.87	0.87	0.87	0.87	0.83	0.83

**Table 6.** Communalities for Different Models.

Item	Unidimensional model		Two-dimensional correlated model		Two-dimensional uncorrelated model				
	R	ChatGPT*	ChatGPT**	R	ChatGPT*	ChatGPT**	R	ChatGPT*	ChatGPT**
Item1	0.31	0.31	0.31	0.69	0.69	0.69	0.69	0.69	0.69
Item2	0.40	0.40	0.40	0.56	0.56	0.56	0.63	0.63	0.63
Item3	0.50	0.50	0.50	0.76	0.76	0.76	0.71,	0.71,	0.71,
Item4	0.32	0.32	0.32	0.53	0.53	0.53	0.45	0.45	0.45
Item5	0.35	0.35	0.35	0.74	0.74	0.74	0.38	0.38	0.38
Item6	0.45	0.45	0.45	0.82	0.82	0.82	0.57	0.57	0.57
Item7	0.39	0.39	0.39	0.65	0.65	0.65	0.55	0.55	0.55
Item8	0.40	0.40	0.40	0.65	0.65	0.65	0.44	0.44	0.44
Item9	0.42	0.42	0.42	0.66	0.66	0.66	0.49	0.49	0.49
Item10	0.34	0.34	0.34	0.78	0.78	0.78	0.54	0.54	0.54



**Figure 4.** Accuracy Estimation Percentage (AEP) and Relative Bias Ratio (RBR) for Factor Loadings.

## Discussion

In recent years, significant progress has been made in data science, paralleling advancements in software development, particularly in artificial intelligence (AI) applications across various domains. Statistical analysis and data science have also benefited from these developments. AI applications can now perform some statistical analyses, and as long as they yield accurate results and adhere to ethical standards, there is no barrier to using AI in data analysis. At this point, the primary concern is whether AI provides accurate findings in data analysis. This study aimed to examine the accuracy of findings obtained through EFA when conducted by AI, specifically in the context of construct validity—a common application in social sciences.

For this purpose, datasets with a sample size of 225, containing 10 items with factor loadings between 0.50 and 0.85, were generated under single-factor, two-factor uncorrelated ( $\phi = 0.00$ ), and two-factor correlated ( $\phi = 0.50$ ) conditions, with 100 repetitions. These datasets were analyzed twice by ChatGPT with a 1-week interval and once using packages in R. To evaluate the findings and performance of R and ChatGPT in EFA, AEP and RBR were used for factor counts and factor loadings. For KMO, total variance explained, and rotated factor loadings, a descriptive comparison was made.

The study found that the KMO values and explained variance ratios from ChatGPT’s initial analysis, a repeated ChatGPT analysis with the same prompt 1 week later, and the R analysis were identical across all conditions. Thus, consistent

results were obtained when ChatGPT analyzed the same dataset with the same prompt at different times. In addition, these findings were consistent with the KMO and explained variance ratios obtained using R packages. KMO and explained variance ratios are calculations directly related to dataset characteristics (Izquierdo et al., 2014; Lloret et al., 2017; Watkins, 2018). From this, it can be concluded that ChatGPT performs accurate calculations for tasks that require purely computational steps. Therefore, researchers can rely on ChatGPT for EFA when given appropriate prompts for these parameters. The findings will likely be accurate under similar data conditions.

One of the most challenging aspects of EFA is determining the number of factors. In this study, the number of factors in the generated datasets was calculated using the empirical Kaiser criterion, the Hull method, and the Kaiser–Guttman criterion. AEPs and RBRs were calculated for these factor counts. For single-factor, two-dimensional uncorrelated, and two-dimensional correlated models, the AEP and RBR of factor counts estimated by ChatGPT (at 1-week intervals) were consistent. These values also aligned with the AEP and RBR of factor counts obtained using R. This suggests that ChatGPT consistently performs well in factor number determination and provides results compatible with R. The AEP and RBR for single-factor structures across two analyses by ChatGPT, conducted 1 week apart, and an R analysis confirmed that the estimated factor counts matched the conditions set in the simulation. Thus, the empirical Kaiser criterion, Hull method, and Kaiser–Guttman criterion methods produce accurate and unbiased results in single-factor models for both R and ChatGPT.

There were minor differences in AEP and RBR between R and ChatGPT's analyses. For example, in single-factor structures, the AEP for the empirical Kaiser criterion was calculated as 93.3 for R, 93.5 for the first analysis by ChatGPT, and 93.4 for the repeated analysis a week later. This difference arose because ChatGPT sometimes failed to execute the entire analysis or specific steps due to misinterpreting the prompt in certain cases. For instance, suppose that in the single-factor structure, ChatGPT did not analyze 3 out of 100 simulated datasets or missed certain calculations in the first round, resulting in 92 datasets reflecting a single-factor structure. In this case, the AEP would be calculated as  $92/(100 - 3) = 0.95$ . If the same datasets were analyzed by ChatGPT 1 week later and all 100 datasets were analyzed without omission, the AEP would be  $92/100=0.92$ . Although both analyses yielded the same outcome for the same number of datasets, the accuracy percentages differed due to minor prompt interpretation issues. These minor differences in AEP and RBR between R and ChatGPT analyses may be attributed to this factor.

In two-dimensional correlated and uncorrelated models, the AEP and RBR between R and ChatGPT's analyses (conducted 1 week apart) were consistent. Therefore, ChatGPT's findings were consistent across different time points and aligned with R's findings. Considering the AEP and RBR, it was observed that estimations for the uncorrelated model diverged from accuracy ( $AEP < 90$ ) and presented biased estimates ( $|RBR| > 0.10$ ). However, it should be noted that this bias originates from the factor determination methods rather than from R or ChatGPT.

The methods used in the study perform well in single-factor structures but may yield biased and low-power estimations in multi-factor structures. Researchers should also consider the theoretical structure when determining factor numbers (Golino et al., 2020; Goretzko & Bühner, 2022; Howard, 2016). Therefore, the factor determination methods used may exhibit bias, especially in complex multi-factor structures, and researchers should be cautious in such cases.

For factor loading evaluations, AEP and RBR values were used. Factor loadings were estimated through EFA, and then compared with the factor loading values defined in the data generation phase to calculate AEP and RBR. The AEP and RBR for factor loadings obtained from analyses conducted by ChatGPT (at 1-week intervals with the same prompt) were consistent. This consistency indicates that ChatGPT can analyze the same dataset with the same prompt at different times and yield the same findings. In addition, these results are consistent with those obtained from R. The RBR for factor loadings was below 0.10, indicating that neither ChatGPT nor R provided biased estimates for factor loadings. Furthermore, the AEP was above 90%, almost reaching 100%, suggesting that the analysis results accurately reflect the conditions manipulated in the simulation. Comparing factor loadings obtained from rotations in two-dimensional models between R and ChatGPT showed identical results, indicating that ChatGPT's performance in factor analysis is compatible with R's results.

The performance of ChatGPT as a data analysis tool was examined through EFA in this study. The findings provide direct insights for EFA and may offer indirect predictions for other analyses. ChatGPT performs well in purely computational tasks or analyses requiring decision-making based on a criterion. This study focused on EFA; further studies could examine its performance in other analyses to verify its accuracy. Further Testing with Complex Models: Future studies could explore ChatGPT's performance with more complex or larger datasets and different types of factor models, such as three-factor or hierarchical structures, to assess scalability and reliability in more sophisticated analyses.

In this study, various factor structures were generated to enable calculations and determinations without the need for researcher judgment. Simulation studies can achieve ideal distributions that are challenging to replicate with real-world data, although real-world data often include various issues (Koçak, 2019). In social sciences, in particular, theoretical constructs and related constructs often need to be considered in research. Typically, when conducting EFA with any statistical software, researchers must consider both the factor loadings and the theoretical structure and dimension that each item measures. Erkuş (2011) noted that merely having the technical skill to use a statistical program or software does not suffice for data analysis. In cases of overlapping items—where an item loads onto multiple factors—the item must be carefully reviewed, sometimes resulting in its removal from the scale (Braeken & Van Assen, 2017; Howard, 2016; Shrestha, 2021). For example, if an item loads onto two different factors, its placement can be determined based on its theoretical structure. Similarly, an item may be positioned in a low-loading factor by

the researcher due to its theoretical significance. Even if an item shows high loading on a factor without overlap, it could be removed if deemed irrelevant to the structure. In these situations, decisions are based on the researcher's judgment rather than statistical outputs alone. ChatGPT's current capabilities involve purely statistical calculations and criterion-based decisions; it does not compare items against the measured structure or theoretical framework. Thus, when analyzing affective, cognitive, or behavioral structures, especially those being defined for the first time, human involvement should not be excluded. While ChatGPT provides accurate results that align with R in statistical terms, caution is advised when interpreting results for complex models or deciding on factor numbers in multidimensional structures, which may require theoretical insight and researcher judgment.

**Enhancing Prompt Clarity:** For optimal results, users should provide clear and detailed prompts when using ChatGPT for statistical analysis. A structured approach to prompting may improve consistency, especially in repeated analyses.

**Integration with Statistical Software:** ChatGPT could benefit from closer integration with established statistical software packages, enhancing its capability to perform specific, advanced statistical analyses while maintaining accuracy.

**User Awareness of Limitations:** Researchers using ChatGPT should remain mindful of its limitations, including potential variability over time and sensitivity to prompt phrasing, and verify results with traditional statistical tools when possible.

Researchers in many fields often face challenges analyzing data from their studies. Researchers whose primary focus is not statistics may feel underqualified, especially in using statistical software. The findings of this study demonstrate ChatGPT's capacity to produce accurate statistical calculations, making it a valuable tool for researchers. Therefore, using ChatGPT in data analysis can be recommended for researchers.

This study was limited to specific conditions and analyses; future research could investigate ChatGPT's performance in other statistical analyses. The following issues are recommended for consideration: (a) enhancing prompt clarity, (b) exploring different parameter settings, (c) assessing the consistency of model outputs, and (d) investigating how AI applications perform in different types of analyses. Other AI Applications and Comparisons: Future studies could also examine the performance of other AI tools, allowing for comparisons across different applications. Since the analyses in this study focused on computations and criterion-based decisions, further research could explore how AI applications perform in analyses that require researcher judgment.

## **Appendix**

### **Data Analysis Process in ChatGPT**

This guide provides a step-by-step explanation of how to effectively use ChatGPT for data analysis. It outlines the steps users should follow when working with ChatGPT,

from accessing the platform to interpreting analysis results. This guide serves as a roadmap for users who are new to ChatGPT or unfamiliar with technical details.

### Step 1: Accessing ChatGPT and Creating an Account

To use ChatGPT, go to [https://openai.com/chatgpt](#). You will need an OpenAI account to access the platform. If you don't have an account, you can register with your email address to create one. After logging in, you'll see a message box on the main screen where you can interact with ChatGPT.

### Step 2: Defining Your Purpose and Analysis Goals

Decide on the purpose of using ChatGPT and the data analysis you intend to conduct. This step helps you define the content of the question or command you want to ask. For example, you may wish to perform data cleaning, exploratory data analysis, or factor analysis. Clarifying your goals will ensure that ChatGPT's analysis outputs align with your needs.

### Step 3: Preparing and Structuring Data

If you plan to conduct data analysis with ChatGPT, ensure that your data is prepared and in an appropriate format. Note that ChatGPT cannot directly process data files, so you'll need to present the data as text or in summary form. For quantitative data, provide it in CSV format. You can share a sample portion of your dataset in tabular format or with key summary information. This will help ChatGPT understand your data and perform the analysis accurately.

### Step 4: Giving ChatGPT Data Analysis Commands

To perform analysis with ChatGPT, type a clear and directive question or command into the message box. A well-defined and clear question or command will enable ChatGPT to conduct a more accurate analysis. For example:

- "Identify the missing values in my dataset and fill them in using an appropriate method."
- "Perform a correlation analysis between the variables in my dataset and display the results."
- "Conduct a factor analysis on this data to determine the number of factors and factor loadings."

Such specific instructions will help ChatGPT understand which analyses to perform. It's helpful to specify the analysis method (e.g., *t*-test, regression, exploratory analysis) you wish to use.

### Step 5: Reviewing and Understanding ChatGPT's Response

After ChatGPT responds to your question, the answer will appear below your message on the same page. At this stage:

- Check if the analysis output provided by ChatGPT meets your purpose.
- Assess the steps included in the response and whether the analysis steps were applied correctly. For example, if you requested factor analysis, verify the factors and loadings provided by ChatGPT.

This process helps you understand the analysis results and the steps ChatGPT followed. If needed, you can request additional details from ChatGPT for clarification.

### Step 6: Asking Follow-up Questions and Deepening Relevant Analysis Steps

You can ask follow-up questions to expand or deepen the analysis process with ChatGPT. After the initial response, you may add more detailed questions or further analysis requests in the message box. For example:

- “Which variables have the highest load values in this factor analysis result?”
- “Can you present this analysis result in a graphical format?”
- “Can you conduct a confidence interval analysis to validate the results?”

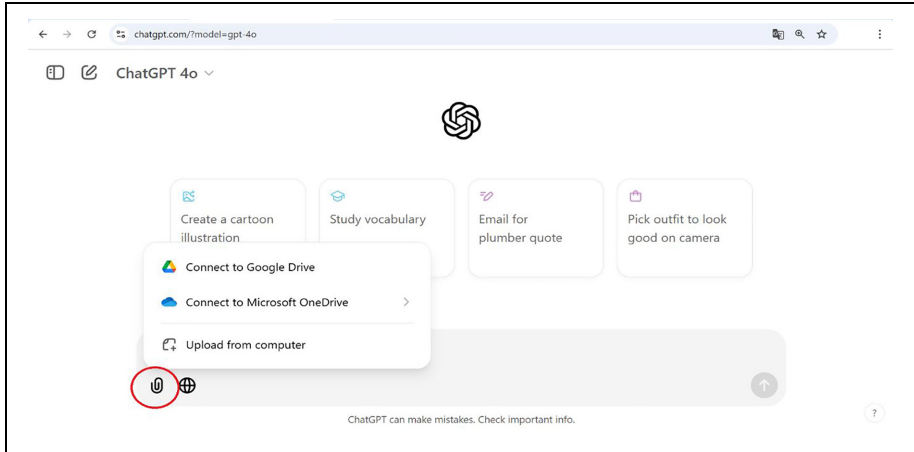
Such follow-up questions enable ChatGPT to elaborate on the analysis and allow you to examine your outputs more thoroughly.

### Step 7: Exploring Alternative Answers—Reproducing the Response

If you find the initial response from ChatGPT insufficient or think a different perspective is needed, you can use the “Regenerate Response” option to receive an alternative answer to the same question. This feature is especially useful for users seeking more explanation or examples in their analysis methods. This way, you can leverage ChatGPT's ability to respond from different angles.

### Step 8: Verifying Analysis Outputs and Interpreting Results

Review the analysis results provided by ChatGPT for reliability and consistency. In statistical analyses, it's essential to verify if the outputs are reasonable. Examine the results and suggestions provided by ChatGPT to ensure they align with your dataset's characteristics. Assess how well the analysis result matches your dataset and research question.



**Figure 5.** ChatGPT Home Screen.

Below, screenshots provide a visual guide on interacting with ChatGPT and navigating the analysis process. These visuals are provided as examples, and users can create similar prompts tailored to their specific analyses.

1. First, log in to ChatGPT.

Once logged in, you will see a chat interface on the screen where you can interact with ChatGPT.

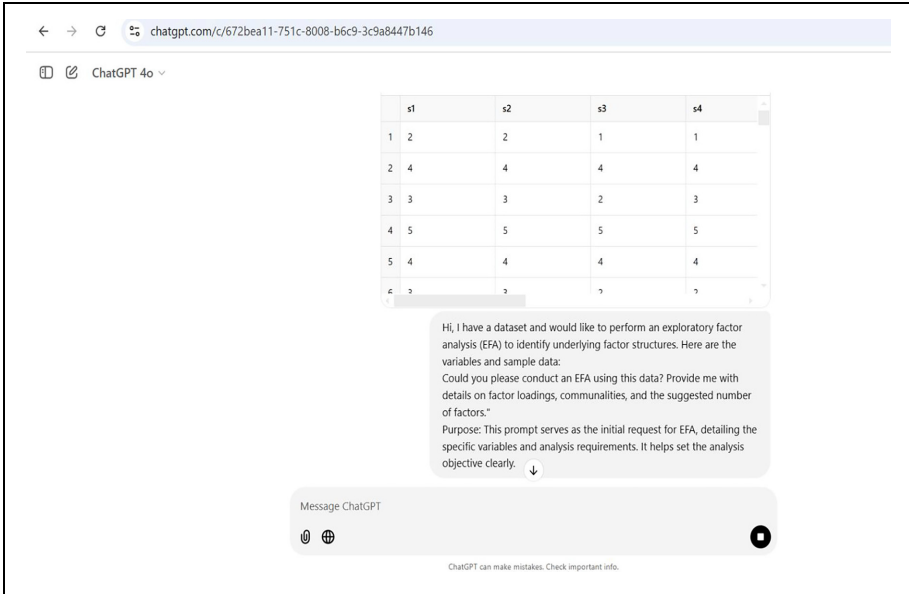
In Figure 5, clicking on the paperclip icon allows you to upload data from your computer or a session directly to ChatGPT. You can write a prompt in the designated area along with your data and send it as a message.

In Figure 6, when the prompt is entered along with the data in the message area, ChatGPT detects the request and begins processing the task.

In Figure 7, when data are provided in an appropriate format and the prompt is clear and sufficiently detailed, ChatGPT executes the identified task. At this point, both the format of the data and the clarity of the prompt are crucial. A clear and precise prompt directly influences the quality of the response. If the initial answer does not meet the user's requirements, additional details or adjustments can be requested to refine the response.

After reviewing ChatGPT's initial response, users may request additional details or clarifications to better meet their needs, as shown in Figure 8. This can include asking for deeper analysis, requesting specific data visualizations, or seeking further explanation of the results.

If necessary, as shown in Figure 9, additional commands and requests can be provided. It's important to note that this step is optional, as ChatGPT may have already produced a response that meets all your requirements in the initial prompt.



**Figure 6.** Sending the First Prompt.

Additional requests, as shown in Figure 10, can be made to further refine or expand the analysis provided by ChatGPT. These may include:

- **Clarifications:** Asking for more detailed explanations of certain parts of the analysis.
- **Further Analysis:** Requesting additional statistical tests or deeper insights based on the initial results.
- **Visualizations:** Asking ChatGPT to provide charts or graphs to better illustrate the findings.
- **Alternative Methods:** Requesting a different approach or methodology to cross-verify the results.

These additional requests allow for a more tailored analysis and can help ensure that the outputs meet specific research or project needs.

The steps of data analysis have been summarized in the figures above. Technically, this process generally involves similar steps, although it may vary depending on data readability, ChatGPT's interpretation of the prompt in alignment with the user's intention, and the level of detail in the request. Ultimately, the process may conclude based on the need for additional requests.

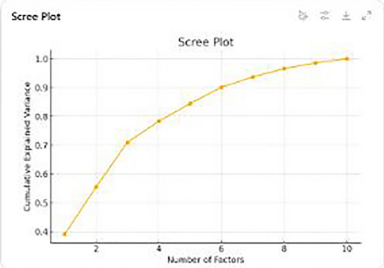
specific variables and analysis requirements. It helps set the analysis objective clearly.

The dataset consists of 10 variables (  $x_1$  through  $x_{10}$  ) with sample values for each. I will proceed with conducting an exploratory factor analysis (EFA) to identify the underlying factor structure, including factor loadings, communalities, and the suggested number of factors.

Let's begin by checking the suitability of the dataset for EFA through measures like the Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity, and then proceed with the factor extraction. [-]

It seems that the required library `factor_analyzer` for calculating KMO and performing Bartlett's test isn't available in this environment. I'll proceed with an alternative approach using available methods to conduct the EFA and analyze the factor loadings.

Let me adjust the process to calculate these values without the missing library. [-]



The screen plot and explained variance analysis suggest that the first three factors account for a significant portion of the variance, with cumulative explained variance reaching approximately 71% by the third factor. This aligns with an initial choice of three factors for further analysis.

I'll now continue by calculating approximate factor loadings for these three factors. [-]

Factor Loadings And Communalities				
		Factor 1	Factor 2	Factor 3
1	$x_1$	-0.37920345425128 61	-0.02826078323450 5135	0.35 1
2	$x_2$	-0.35639635062344 19	0.093212735122344 11	0.28 64
3	$x_3$	-0.37794888656996 206	-0.141822597196019 69	0.27 64

I've provided the factor loadings and communalities for the first three factors based on this dataset. These results should help in interpreting the underlying factor structure, and the communalities indicate the portion of variance in each variable explained by the extracted factors. Let me know if you need further analysis or adjustments. [-]

Is this conversation helping so far?  
 Can you interpret the factor loadings?    Could we try with four factors?

Message ChatGPT

ChatGPT can make mistakes. Check important info.

Figure 7. ChatGPT's Response to the First Prompt.

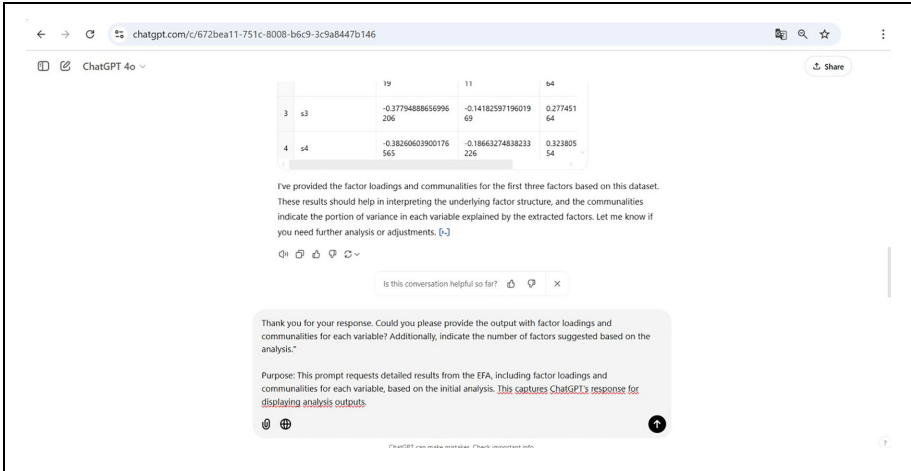


Figure 8. Details Requested Based on the Initial Response.

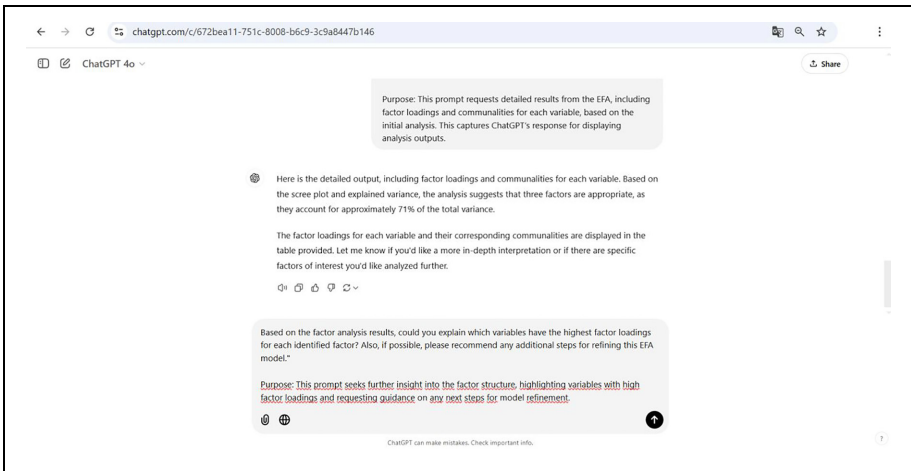
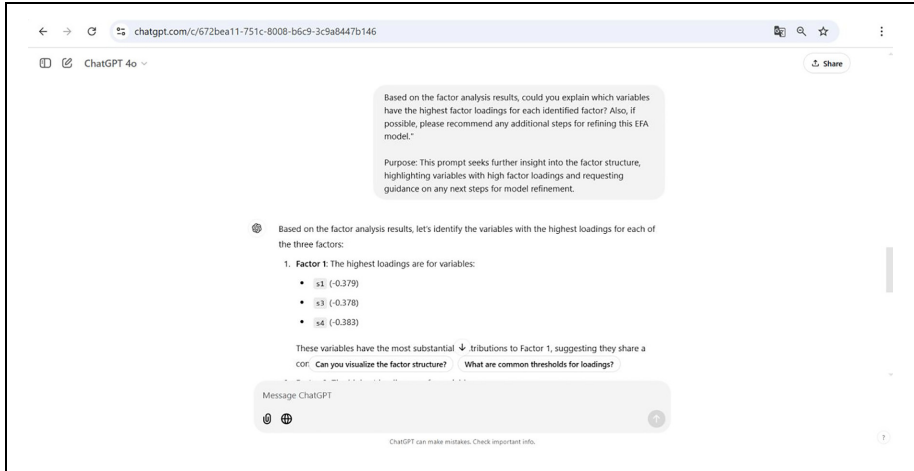


Figure 9. ChatGPT's Response to the Request for Details.



**Figure 10.** Additional Requests.


## Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Duygu Koçak  <https://orcid.org/0000-0003-3211-0426>

## References

- Adelson, J. L., & McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-type scale. *Educational and Psychological Measurement, 70*(5), 796–807. <https://doi.org/10.1177/0013164410366694>
- Alshurafat, H. (2023, February 2). The usefulness and challenges of chatbots for accounting professionals: Application on ChatGPT. *SSRN*. <https://ssrn.com/abstract=4345921>
- Ambati, S. H., Ridley, N., Branca, E., & Stakhanova, N. (2024, September 2–4). Navigating (in)security of AI-generated code. In *2024 IEEE international conference on cyber security and resilience (CSR)* (pp. 1–8). IEEE.
- Baidoo-Anu, D., & Owusu Ansah, L. (2023, January 25). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN*. <https://ssrn.com/abstract=4337484>

- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186–203. [https://doi.org/10.1207/s15328007sem1302\\_2](https://doi.org/10.1207/s15328007sem1302_2)
- Bernaards, C., Jennrich, R., & Gilbert, M. P. (2015). *Package “gparotation.”* <https://cran.r-project.org/web/packages/GPARotation/index.html>
- Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, 22(3), 450–466. <https://doi.org/10.1037/met0000074>
- Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., & . . . De Rosario, M. H. (2018). *Package “pwr”* (R package version, 1(2)). <https://doi.org/10.32614/CRAN.package.pwr>
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351.
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Pearson.
- Cribben, I., & Zeinali, Y. (2023). *The benefits and limitations of ChatGPT in business education and research: A focus on management science, operations management and data analytics*. <https://ssrn.com/abstract=4390951>
- Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.
- Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21, 17–23. <https://doi.org/10.3354/ese00195>
- Dinno, A., & Dinno, M. A. (2018). *Package “paran”* (R package version 1(2)). <https://doi.org/10.32614/CRAN.package.paran>
- Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829–848). Blackwell.
- Epskamp, S. (2017). *Package “semPlot.”* <https://cran.r-project.org/web/packages/semPlot/semPlot.pdf>
- Erkuş, A. (2011). *Davranış bilimleri için bilimsel araştırma süreci*. Seçkin Yayıncılık.
- Field, A. (2009). *Discovering statistics using SPSS*. Sage.
- 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>
- Frantz, M. E. (2024). *Measurement and development for automated secure coding solutions*. (Doctoral dissertation, Virginia Polytechnic Institute and State University). Blacksburg, VA. <https://techworks.lib.vt.edu/items/017ec015-79b0-40d3-8f4c-c992f743b540>
- Gillani, N., Eynon, R., Chiabaut, C., & Finkel, K. (2023). Unpacking the “black box” of AI in education. *Educational Technology & Society*, 26(1), 99–111.
- Golino, H. F., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., Thiagarajan, J. A., & Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, 25(3), 292–320. <https://doi.org/10.1037/met0000255>
- Goretzko, D., & Bühner, M. (2022). Factor retention using machine learning with ordinal data. *Applied Psychological Measurement*, 46(5), 406–421.

- Harwell, D. E., Mortimer, M. D., Knobler, C. B., Anet, F. A., & Hawthorne, M. F. (1996). Auracboranes with and without Au-Au Interactions: An Unusually Strong Auophilic Interaction. *Journal of the American Chemical Society*, *118*(11), 2679–2685. <https://doi.org/10.1021/ja953976y>
- Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388. <https://doi.org/10.1177/0013164414548576>
- Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62. <https://doi.org/10.1080/10447318.2015.1087664>
- Ifthikhar, L. (2023). DocGPT: Impact of ChatGPT-3 on health services as a virtual doctor. *EC Paediatrics*, *12*(1), 45–55.
- Izquierdo, I., Olea, J., & Abad, F. J. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema*, *26*(4), 395–400. <https://doi.org/10.7334/psicothema2013.349>
- Ş Kalaycı. (Ed.). (2010). *SPSS uygulamalı çok değişkenli istatistik teknikleri*. Asil Yayın Dağıtım.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kelley, K., Kelley, M. K., & Imports, M. A. S. S. (2018). *The MBESS R package* [Computer software]. <https://CRAN.R-project.org/package=MBESS>
- Koçak, D. (2019). A method to increase the power of Monte Carlo method: Increasing the number of iteration. *Pedagogical Research*, *5*(1), Article em0049.
- Latendresse, J., Khatoonabadi, S., Abdellatif, A., & Shihab, E. (2024). Is ChatGPT a good software librarian? An exploratory study on the use of ChatGPT for software library recommendations. arXiv:2408.05128.
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Liang, H., Li, X., Xiao, D., Liu, J., Zhou, Y., Wang, A., & Li, J. (2023). Generative pre-trained transformer-based reinforcement learning for testing web application firewalls. *IEEE Transactions on Dependable and Secure Computing*, *21*(1), 309–324. <https://doi.org/10.1109/TDSC.2022.3158419>
- Lloret, S., Ferreres, A., Hernandez, A., & Tomas, I. (2017). The exploratory factor analysis of items: Guided analysis based on empirical data and software. *Anales de Psicología*, *33*(2), 417–432. <https://doi.org/10.6018/analesps.33.2.270211>
- McGee, R. W. (2023, February 15). Annie Chan: Three short stories written with ChatGPT. *SSRN*. <https://ssrn.com/abstract=4359403>
- Mellor, D., & Moore, K. A. (2014). The use of Likert scales with children. *Journal of Pediatric Psychology*, *39*(3), 369–379. <https://doi.org/10.1093/jpepsy/jst079>
- Mhlanga, D. (2023, February 11). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. *SSRN*. <https://ssrn.com/abstract=4354422>
- Mollick, E. R., & Mollick, L. (2022, December 13). New modes of learning enabled by AI chatbots: Three methods and assignments. *SSRN*. <https://ssrn.com/abstract=4300783>

- Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, *10*(2), 60–70. <https://doi.org/10.1027/1614-2241/a000068>
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620.
- Navarro-Gonzalez, D., Lorenzo-Seva, U., & Navarro-Gonzalez, M. D. (2020). Package “EFA.MRFA.” <https://doi.org/10.32614/CRAN.package.EFA.MRFA>
- Nemorin, S., Vlachidis, A., Ayerakwa, H. M., & Andriotis, P. (2023). AI hyped? A horizon scan of discourse on artificial intelligence in education (AIED) and development. *Learning, Media and Technology*, *48*(1), 38–51. <https://doi.org/10.1080/17439884.2022.2147831>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows*. Open University Press.
- Raiche, G., Magis, D., & Raiche, M. G. (2020). Package “nFactors” [Repository CRAN]. <https://doi.org/10.32614/CRAN.package.nFactors>
- R Development Core Team. (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org>
- Revelle, W., & Revelle, M. W. (2015). Package “psych.” *The Comprehensive R Archive Network*, *337*(338), 161–165. <https://CRAN.R-project.org/package=psych>
- Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package “mass.” *Cran R*, *538*, 113–120.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.
- Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., & . . . Jorgensen, T. (2017). Package “lavaan.” <https://doi.org/10.32614/CRAN.package.lavaan>
- Rudolph, J., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), 342–363.
- Ryan, E. G., Vitoratou, S., Goldsmith, K. A., & Chalder, T. (2018). Psychometric properties and factor structure of a long and shortened version of the cognitive and behavioural responses questionnaire. *Psychosomatic Medicine*, *80*(2), 230–237. <https://doi.org/10.1097/PSY.0000000000000536>
- Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4–11. <https://doi.org/10.12691/ajams-9-1-2>
- Singh, B., Kumar, R., & Singh, V. P. (2022). Reinforcement learning in robotic applications: A comprehensive survey. *Artificial Intelligence Review*, *55*, 945–990. <https://doi.org/10.1007/s10462-022-10166-7>
- Siow, J. Y. M., Chan, A., Østbye, T., Cheng, G. H.-L., & Malhotra, R. (2017). Validity and reliability of the positive aspects of caregiving (PAC) scale and development of its shorter version (S-PAC) among family caregivers of older adults. *The Gerontologist*, *57*, e75–e84. <https://doi.org/10.1093/geront/gnw198>
- Tan, Ş. (2016). *SPSS ve Excel uygulamalı temel istatistik-I*. Pegem Akademi Yayıncılık.
- Tang, H., Mao, L., Wang, F., & Zhang, H. (2022). A validation study for a short-version scale to assess 21st century skills in flipped EFL classrooms. *Oxford Review of Education*, *48*(2), 148–165. <https://doi.org/10.1080/03054985.2021.1935226>
- Tassoti, S. (2024). Assessment of students use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education. *Journal of Chemical Education*, *101*, 2475–2482.

- Terwiesch, C. (2023). *Would ChatGPT get a Wharton MBA? A prediction based on its performance in the operations management course*. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania. <https://mackinstitute.wharton.upenn.edu/2023/terwiesch-chatgpt-wharton-mba/>
- Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., & . . . Ognibene, D. (2023, November). Learning to prompt in the classroom to understand AI limits: A pilot study. In *International conference of the Italian association for artificial intelligence* (pp. 481–496). Springer Nature.
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1), 1–24. <https://doi.org/10.1186/s40561-023-00205-7>
- Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, 44(3), 219–246. <https://doi.org/10.1177/0095798418771807>
- Wood, D. A., Achhpilia, M. P., Adams, M. T., Aghazadeh, S., Akinyele, K., Akpan, M., & Kuruppu, C. (2023). The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education*, 38, 1–28. <https://doi.org/10.2308/issues-2023-005>