JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v13i1.34173
http://journal.uinjkt.ac.id/index.php/jp3i

Optimal Scale Points for Reliable Measurements: Exploring the Impact of Scale Point Variation

Raoda Ismail [1,2], Heri Retnawati [1], Sugiman [1], Farida Agus Setiawati [1], Okky Riswandha Imawan [1,2] , Purwoko Haryadi Santoso [1]

Educational Research and Evaluation, Graduated School, Yogyakarta State University, Indonesia[1]
Mathematics Education, Cenderawasih University, Indonesia[2]

raodaismail.2021@student.uny.ac.id

Abstract

Ensuring reliable measurements is crucial for minimising errors in assessments. The assessment community commonly employs the evaluation of reliability coefficients to estimate the dependability of test scores. Despite its significance, limited research has explored the relationship between the estimated reliability coefficient and the number of scale points utilised. This study aims to provide valuable insights to practitioners by investigating the optimal number of scale points required for the most accurate reliability coefficient estimation. Using simulated data, the research scrutinises scales with varying points, ranging from 2 to 11. The results reveal a substantial impact of the number of scale points on reliability estimation. The most accurate estimate of reliability is obtained for scales with 8 points. This study helps us understand the optimal number of scale points for reliable measurements and guides future assessment improvements.

Keywords: number of scale points, non-normal, reliability coefficient.

Abstrak

Memastikan pengukuran yang dapat diandalkan sangat penting untuk meminimalkan kesalahan dalam penilaian. Komunitas penilaian umumnya menggunakan evaluasi koefisien reliabilitas untuk memperkirakan keterandalan skor tes. Meskipun memiliki signifikansi, penelitian yang terbatas telah menjelajahi hubungan antara koefisien keandalan yang diestimasi dan jumlah poin skala yang digunakan. Penelitian ini bertujuan memberikan wawasan berharga kepada praktisi dengan menyelidiki jumlah poin skala yang optimal yang diperlukan untuk estimasi koefisien keandalan paling akurat. Dengan menggunakan data simulasi, penelitian ini memeriksa skala dengan jumlah poin yang bervariasi, mulai dari 2 hingga 11. Hasil penelitian menunjukkan dampak signifikan dari jumlah poin skala terhadap estimasi keandalan. Estimasi keandalan paling akurat diperoleh untuk skala dengan 8 poin. Penelitian ini tidak hanya membantu kita memahami jumlah poin skala yang optimal untuk pengukuran yang dapat diandalkan tetapi juga membimbing perbaikan masa depan dalam penilaian.

***Kata kunci**: jumlah titik skala, ketidaknormalan, koefisien reliabilitas.*

Introduction

Understanding the reliability of test scores is crucial for assessing consistency and dependability. Reliability reflects how closely observed scores match actual scores (Allen & Yen, 1979; Retnawati, 2020). It ensures a test yields consistent results across different administrations or item sets. Reliability measures the consistency of a test's measurements (Ebel & Frisbie, 1991). This study aims to guide practitioners in determining the optimal number of scale points for accurate reliability estimation. More scale points are expected to enhance consistency and reveal biases in reliability coefficient estimation results (Retnawati, 2020). Analysing both scale points and distribution helps uncover any biases in reliability coefficients.

Notably, many educational and psychological assessments utilise instruments with a five-point scale. Previous research has indicated no significant difference in reliability estimates for scores based on four-point and six-point scales (McColly & Remstad, 1965). However, subsequent research has shown that increasing the number of scale points can enhance the estimated reliability coefficient up to a certain point, particularly when the scale does not exceed fifteen points (Shumate et al., 2007). Several subsequent studies have also demonstrated that the length of the scale can influence resulting reliability estimates (Alan & Kabasakal, 2020; Raadt et al., 2021; Rahayu & Abidin, 2017). Moreover, research suggests that an eleven-point Likert scale can increase generalizability (H. Wu & Leung, 2017).

In the domains of education and psychology, it is imperative to consider not only the impact of the number of scale points on the estimation of the reliability coefficient but also to examine and identify the influence of score distribution. This involves understanding how the distribution of scores can affect the estimation process, with scores often exhibiting skewness or a statistical distribution where the value of excess kurtosis is negative (platykurtic) (Tsai et al., 2017; X. Z. Wu, 2020). Tsai (2017) proposed a new family of hyperbolic power transformations to improve the normality of raw data with varying degrees of slope and kurtosis. This new family proves effective in converting the distribution of platykurtic or bimodal data to normal. The proposed transformation family is illustrated through a simulation study and real examples of data on mathematics achievement test scores.

This simulation-based study builds upon prior research by systematically examining how the number of scale points and score distribution can impact the bias in estimating the reliability coefficient. Through the generation of data for various numbers of scale points, ranging from a two-point scale to an eleven-point scale, this study systematically analyses the resulting reliability coefficients. The ultimate objective of this scientific endeavour is to enhance measurement precision, a fundamental aspect of scientific advancement (Greco et al., 2018). The chosen indicator for measuring precision is the reliability coefficient, focusing specifically on the Cronbach Alpha reliability coefficient. Cronbach's alpha is widely used in the social and organisational sciences (Bonett & Wright, 2015) and is the most commonly used method for estimating internal consistency reliability. Despite its popularity, it has limitations, and alternative measures, such as omega coefficients, which are especially beneficial for applied research, have been proposed (Trizano-Hermosilla & Alvarado, 2016).

Due to the influence of the number of items and the presence of parallel items on the alpha coefficient, scale developers sometimes mistakenly include an excessive number of items, assuming that a high alpha value indicates a reliable psychometric scale. However, a high alpha value may signify item redundancy, where numerous items have weak connections to the underlying construct (Panayides, 2013). Sijtsma's comprehensive investigation of the utility of the Alpha coefficient has demonstrated that employing multiple-item tests is the most effective approach (2009). An exceedingly high alpha value may suggest an excessively long scale, including parallel items, or a narrow scope of constructs being considered (Panayides, 2013). Consequently, research is scarce focusing on maximising reliability in rubrics that incorporate multiple scale choices.

Hence, the primary research inquiries are as follows: What is the optimal number of scale points that yields the most accurate estimation of the reliability coefficient? To address these research questions, the reliability coefficients will be examined through a combination of simulation and empirical data

analysis for each scale point, ranging from 2 to 11. These various scale points complement one another, facilitating the use of analytical methods to elucidate the relationships between them. This study endeavours to illuminate the resulting reliability scenarios using simulation data and empirical observations.

Methods

A practical-based method emphasises applying learned concepts in real-world scenarios, combining theory with real-world actions to enable researchers or practitioners to observe, understand, and apply strategies effectively. Moreover, this approach is oriented towards direct implementation in real-life situations or utilisation in everyday life.

To perform the simulation, the researcher utilised Microsoft Excel through the Random Number Generation and Sampling menu. The first step involved generating ten sets of data that follow a normal distribution, containing random numbers with sample sizes of 200 and 1000. The generated data was then tested to assess whether it adheres to the normality assumption. The importance of normal distribution cannot be ignored as it serves as a fundamental assumption in numerous statistical procedures. Violating the normality assumption may render interpretations and inferences unreliable or invalid. Three common procedures for assessing whether a random sample of independent observations of size n is derived from a population with a normal distribution include graphical methods (histogram, Q-Q-plot), numerical methods (skewness and kurtosis indices), and formal normality tests (such as Kolmogorov-Smirnov (KS) test, Anderson-Darling (AD) test, and Shapiro-Wilk (SW) test). This study will analyse it using graphical methods and formal normality tests.

In many studies on the construction and validation of psychometric scales, a heavy emphasis is placed on the alpha coefficient (Cronbach, 1951). The following equation gives the formula for alpha:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right)$$

Where n is the number of items, Vt is the variance of the total score, and Vt is the variance of the item scores. Cronbach (1951) describes alpha as a generalisation of the Kuder-Richardson equivalence coefficient (K-R20), which has the following essential properties: (a) α is the mean of all possible split-half coefficients (b) α is the expected value when two random samples of items from the set as in a given test is correlated (c) α is the lower bound of the coefficient of precision, (d) α is estimated, and is the lower bound for the proportion of test variance attributable to common factors among items, (e) α is the upper bound concentration in the first-factor test among items.

Guttman's lambda-2 (λ2) and its role as an estimate of reliability are part of a series of six lambda values proposed by Guttman in 1945, with lambda-2 being the second in this series. Alongside lambda-3 (equivalent to Cronbach's alpha), lambda-2 is commonly used in psychometric analyses. It shares similarities with Cronbach's alpha as both represent correlation estimates between scores for parallel measures. When comparing lambda-2 to alpha, the former is considered more substantial when dealing with combined tasks, as it measures the correlation between scores for parallel measures. In contrast, alpha measures the correlation between scores for randomly equivalent measures. Despite its complex formula involving item covariances, Guttman's lambda-2 consistently equals or surpasses alpha for tests, adding value as a reliability estimate in psychometric analysis. The formula for Guttman's λ2 is given by the following equation (Widhiarso & Mardapi, 2011):

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{2k}{k-1}\sum\sum v^2{}_{Y1Y2}}}{S^2{}_X}$$

http://journal.uinjkt.ac.id/index.php/jp3i

In psychometrics, internal consistency reliability of tests has been a primary concern, with Cronbach's alpha (α) being a commonly used estimation. Despite its widespread use due to ease of calculation, research suggests that alpha can underestimate test reliability and overestimate the first-factor saturation. Alpha is equivalent to Guttman's Lambda 3 (λ3), which can be derived using specific mathematical formulas. However, alpha's limitations become evident when tests exhibit microstructures or "lumpiness," where coefficients like beta and omega_hierarchical offer more accurate alternatives. On the other hand, Guttman's Lambda 6 (λ6) evaluates how much variance in each item can be explained by other items, introducing a new dimension to reliability measurement. Λ6 is also sensitive to the "lumpiness" in tests and provides different estimates than alpha depending on the test's structural conditions, adding complexity in selecting appropriate methods for measuring internal consistency in a test. While alpha remains a commonly reported metric for its ease of computation, understanding the weaknesses and alternatives such as λ6, beta, and omega_hierarchical adds nuance to the complex landscape of measuring test reliability. Further research is needed to refine these approaches and enhance the accuracy of reliability measurements in psychometrics. The formula for Guttman's λ2 is given by the following equation (Widhiarso & Mardapi, 2011):

$$\lambda_6 = 1 - \frac{\sum e^2{}_j}{\sigma^2{}_X}$$

The dataset encompasses a comprehensive collection of information, with 20 individual items corresponding to each distinct scale point, spanning a range from 2 to 11 on the measurement scale. To ensure a thorough analysis, two different sample sizes were utilised, specifically consisting of 200 and 1000 data points, allowing for a more robust exploration of the generated data. All analyses were conducted using JASP software.

Results and Discussion

Results

The research outcomes are presented in four distinct tables: the Descriptives Table, the Fit Statistics Table, the Q-Q Plots Histogram Table, and the Frequentist Scale Reliability Statistics Table. All analyses were conducted using JASP software.

Descriptives

The Descriptives Table specifically elucidates mean and standard deviation values, which are meticulously presented in Table 1. Statistically, it is stated that a larger sample size is expected to yield increasingly better results. With a large sample, the mean and standard deviation obtained are highly likely to resemble the population mean and standard deviation. This is because the sample size is related to the testing of statistical hypotheses. Although a larger sample would be better, a small randomly selected sample could also accurately reflect the population. The standard deviation value determines the data spread in a sample and how close the data points are to the mean value. The standard deviation reflects the average deviation of data from the mean. Standard deviation can depict the extent of data variation, where if the standard deviation value is greater than the mean value, it means the mean value is a poor representation of the entire data set. However, if the standard deviation value is smaller than the mean value, it indicates that the mean value can represent the entire data set. Table 1 shows that all standard deviation values are close to the mean, thus indicating that the mean value can represent the entire data set.

Table 1. Descriptives

Number of Scale Points	Mean		Std. Deviation	
	N=200	N=1000	N=200	N=1000
2	30.060	29.845	2.348	2.178
3	40.365	40.157	3.304	3.702
4	50.505	49.930	4.880	5.011
5	60.635	59.976	6.363	6.193
6	69.520	69.812	7.069	7.736
7	79.290	78.375	9.381	9.721
8	88.375	98.375	10.742	10.721
9	100.235	100.195	11.479	11.843
10	108.195	107.990	12.962	13.205
11	118.950	118.950	12.063	14.035

From the given table, it can be observed that there is data for two conditions, namely N=200 and N=1000, for the mean and standard deviation across various scale numbers. In general, there is a pattern that with an increase in the number of scales, both the mean and standard deviation tend to increase. This is reasonable because, with more scales, the data has a more significant potential for variability. However, it is worth noting that there are some significant differences between the results for N=200 and N=1000 at certain points. For example, at scale 8, there is a considerable difference in the mean between the two conditions (88.375 vs 98.375) and the standard deviation (10.742 vs 10.721). This indicates a significant deviation in the data between these two conditions.

Furthermore, at scale 11, there is a considerable difference in the standard deviation between the two conditions (12.063 vs 14.035), while the mean remains the same. This suggests significant variation in the data distribution between the two conditions. Overall, although at some points the data are quite close, there are significant deviations in the data between the two conditions. This indicates that other factors may influence data distribution between the two conditions that need further consideration.

Fit Statistics

Table 2, the Fit Statistics table, is crucial as it provides detailed information on various statistical values and p-values associated with reliability coefficients, including the Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk coefficients. These coefficients are essential indicators of the fit between observed data and theoretical distributions, which are critical in assessing the reliability and validity of statistical models.

From Table 2, researchers and analysts can obtain valuable insights into the goodness of fit of their statistical models. Specifically, they can assess how well the observed data align with the theoretical distributions assumed by the models. The statistical values and p-values provided in the table allow researchers to determine whether the observed data significantly deviate from the expected distributions.

The criteria for evaluating the fit statistics typically involve comparing the obtained statistical values and p-values with predetermined thresholds or benchmarks. For example, smaller p-values indicate a poorer fit between the observed data and theoretical distributions, while more significant p-values suggest a better fit. Similarly, larger statistical values indicate more significant discrepancies between observed and expected distributions.

By examining Table 2, readers can discern which statistical values and p-values fall within acceptable ranges and which deviate significantly. This information is crucial for making informed decisions about statistical models' reliability and validity and identifying areas that may require further investigation or adjustment. In summary, Table 2 plays a vital role in assessing the goodness of fit of statistical models by providing detailed information on various fit statistics and p-values. It enables researchers to evaluate their models' reliability and validity and identify areas for improvement or further investigation.

Table 2. Fit Statistics

Scale	Kolmogorov-Smirnov				Anderson Darling				Shapiro-Wilk			
	200		1000		200		1000		200		1000	
	Stat	p	Stat	p	Stat	p	Stat	p	Stat	p	Stat	p
2	.107	.020	.100	< .001	**2.133**	**.078**	**8.823**	**< .001**	2.133	.078	.981	< .001
3	.067	.323	.069	< .001	.913	.406	3.662	.013	.913	.406	.992	< .001
4	.056	.556	.052	.009	.593	.654	1.878	.107	.593	.654	.995	.003
5	.063	.397	.048	.021	.471	.776	1.560	.163	.471	.776	.996	.012
6	.058	.516	.058	.003	.366	.892	1.774	.123	.366	.892	.994	< .001
7	.050	.701	.052	.009	.487	.760	2.820	.034	.487	.760	.979	< .001
8	.052	.646	.052	.009	.564	.682	2.820	.034	.564	.682	.979	< .001
9	.047	.773	.034	.209	.297	.940	.568	.678	.297	.940	.998	.227
10	.040	.903	.057	.003	**.190**	**.993**	**2.301**	**.063**	.190	.993	.989	< .001
11	.035	.970	.035	.183	.209	.988	1.044	.335	.209	.988	.996	.012

The table shows data for two conditions, N=200 and N=1000, related to the Kolmogorov-Smirnov, Anderson Darling, and Shapiro-Wilk statistics values across various scales. In evaluating the fit statistics, the main focus is on the statistics values and p-values. These statistics values reflect the extent of deviation between the observed data and the assumed theoretical distribution. Meanwhile, the p-values indicate the statistical significance of such deviations. Generally, the smaller the statistics values and the larger the p-values, the better the fit between the observed data and the theoretical distribution. Conversely, if the statistics values are large and the p-values are small, it indicates a significant deviation between the observed data and the theoretical distribution.

The table results show variation in the statistics values and p-values between N=200 and N=1000 across some scales. For instance, at scale 2, the p-value for Anderson-Darling at N=1000 is significantly smaller than at N=200, indicating a significant deviation at N=200. However, for other scales, such as scale 10, there is no significant difference between the two conditions. Therefore, it can be concluded that not all data are in accordance with the fit statistics. There are significant deviations across some scales, especially at N=200. Hence, further analysis is required to understand the factors causing these deviations, and adjustments may be needed for the statistical model used.

Histogram and Q-Q Plot

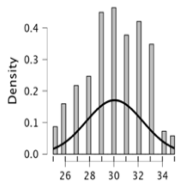
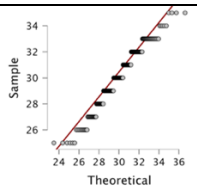
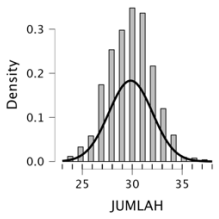
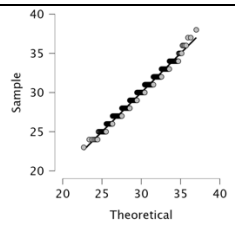
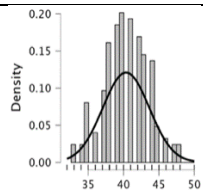
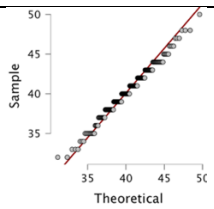
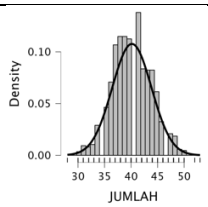
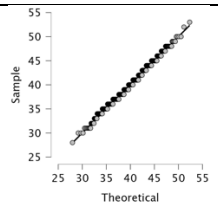
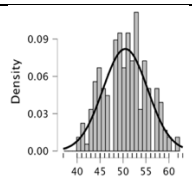
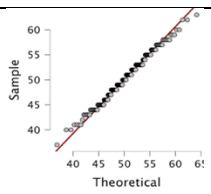
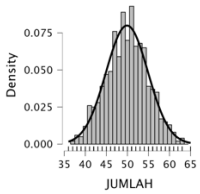
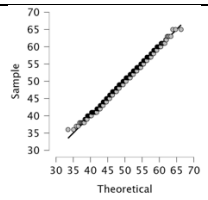
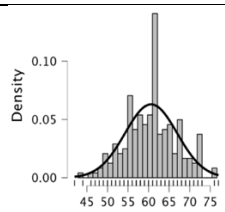
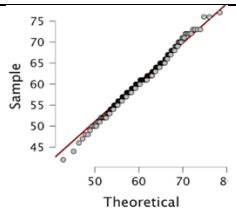
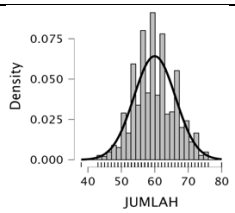
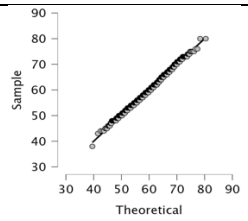
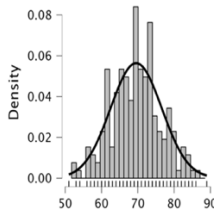
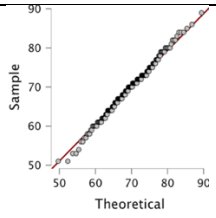
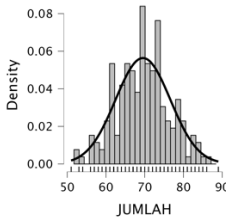
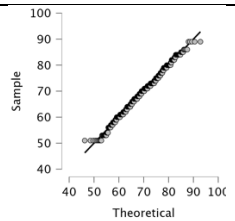
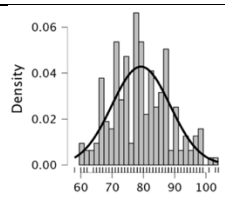
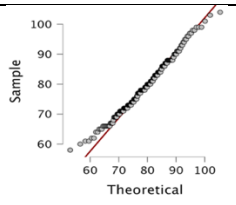
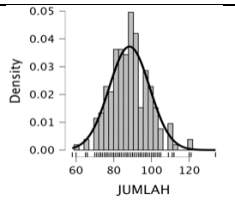
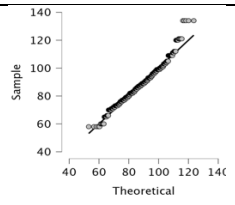
Table 3 presents detailed representations of the Histogram and Quantile-Quantile (Q-Q) Plot for each specified number of scale points, categorised under N=200 and N=1000. The Histogram vs Theoretical PDF comparison allows for an assessment of how closely the observed data matches the theoretical probability density function (PDF) represented by the histogram. Meanwhile, the Q-Q Plot provides a graphical comparison between the quantiles of the observed data and the quantiles of a theoretical distribution, aiding in evaluating the similarity between the two datasets.

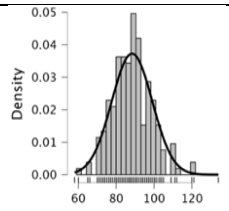
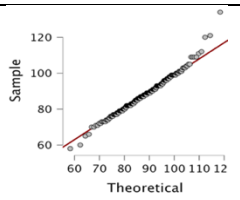
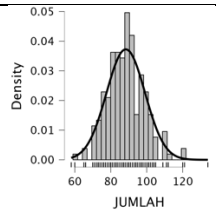
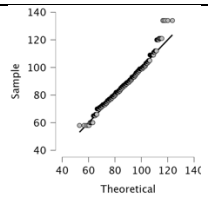
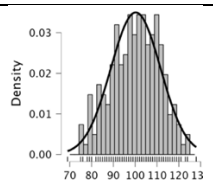
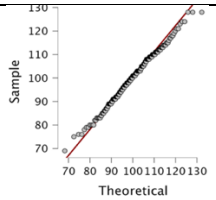
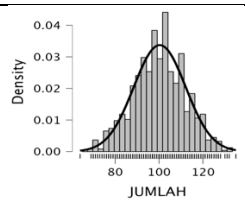
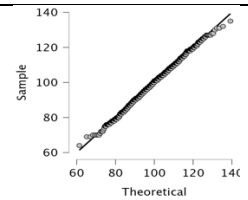
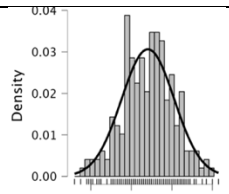
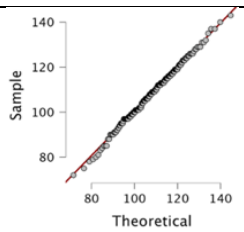
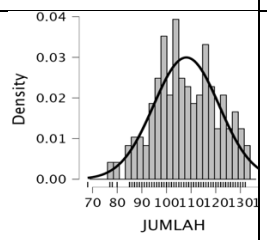
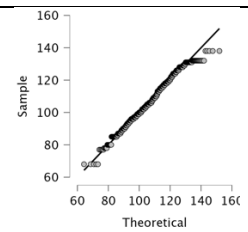
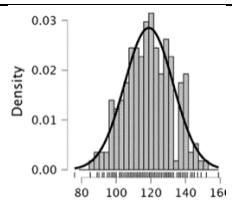
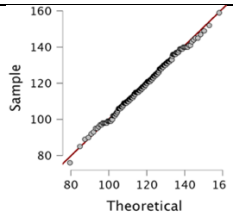
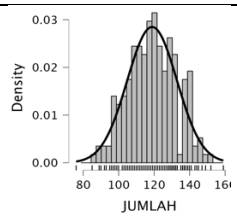
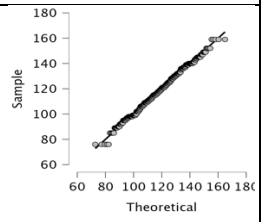
Each entry in the table corresponds to a specific scale point. For instance, under Scale 2, there should be representations of the Histogram vs. Theoretical PDF and Q-Q Plot for both N=200 and N=1000. One would need to closely examine the representations provided in the table to assess whether there is

any deviation or all data aligns appropriately. If the histograms closely resemble the theoretical PDF and the points in the Q-Q Plot lie along a straight line, the observed data matches the theoretical distribution well. On the other hand, significant deviations between the histograms and theoretical PDFs, or non-linear patterns in the Q-Q Plots, would indicate discrepancies between the observed and theoretical distributions.

Therefore, by analysing the representations provided in Table 3, one can determine whether deviations exist between the observed data and the theoretical distributions for each specified number of scale points.

Table 3. Histogram and Q-Q Plot

Scale	N=200		N=1000	
	Histogram vs. Theoretical PDF	**Q-Q Plot**	**Histogram vs. Theoretical PDF**	**Q-Q Plot**
2				
3				
4				
5				
6				
7				

Scale	N=200		N=1000	
	Histogram vs. Theoretical PDF	Q-Q Plot	Histogram vs. Theoretical PDF	Q-Q Plot
8				
9				
10				
11				

At scale 2, for N=200 and N=1000, the results indicate that the histogram closely does not resemble the theoretical PDF. This suggests that the distribution of the observed data does not match the expected theoretical distribution. Additionally, the points in the Q-Q Plot are not quite aligned with the straight line, indicating that the quantiles of the observed data do not consistently match the quantiles of the theoretical distribution. A similar situation occurs at scale 3 for N=200, where the histogram does not closely resemble the theoretical PDF, and the points in the Q-Q Plot are not aligned with the straight line. This suggests that the distribution of data at scale three and N=200 also does not fit the expected theoretical distribution well.

These findings indicate a mismatch between the observed data and the theoretical distribution at scale 2 and scale 3, especially for N=200. This suggests the need to review further the statistical model used or reassess the data collection process.

Frequentist Scale Reliability Statistics

Table 4, "Frequentist Scale Reliability Statistics," presents the reliability coefficients for each estimate, namely Cronbach's α, Guttman's $\lambda 2$, and Guttman's $\lambda 6$, for specific scales. The data is divided based on two conditions, namely N=200 and N=1000. In evaluating the scale's reliability, the main focus is on the coefficients' values in the table. Cronbach's α, Guttman's λ2, and Guttman's λ6 are essential indicators of consistency in scale measurement. These values range from -1 to 1, where higher values indicate higher reliability.

Table 4. Frequentist Scale Reliability Statistics

Number of Scale Points	Cronbach's α	Guttman' s λ2	Guttman' s λ6	Cronbach's α	Guttman' s λ2	Guttman' s λ6
N	200	200	200	1000	1000	1000
2	**.100**	**.165**	**.193**	-.056	-.023	-.037
3	-.217	-.125	-.097	.056	-.023	-.037
4	-.052	.024	.054	.009	.040	.029
5	.029	.086	.098	.038	-.004	-.018
6	-.142	-.055	-.035	**.142**	-.055	-.035
7	-.097	-.014	.014	.039	-.005	-.018
8	**.095**	**.154**	**.172**	**.097**	**.156**	**.174**
9	.004	.081	.110	.048	.081	.069
10	.004	.081	.110	.077	**.137**	**.164**
11	-.032	.041	.065	-.032	.041	.065

From the table results, there is variation in the reliability coefficient values between N=200 and N=1000 across various numbers of scales. There are significant differences in these values at some points. For example, at scale 2, there is a considerable difference between the reliability coefficient values for N=200 and N=1000. This indicates a deviation in the scale's reliability between the two conditions. Similarly, there is significant variation between the two conditions at some other scales. Therefore, it can be concluded that not all data align with the measured scale reliability. Significant deviations exist in some reliability coefficient values between the N=200 and N=1000 conditions. This suggests the need for further review to understand the factors causing differences in scale reliability.

The results indicate several important patterns: First, there is a positive relationship between the increase in the number of scale points and the increase in the resulting variance. This means that the more scales used, the greater the variability of the resulting data. Second, as the number of scale points increases, the distribution of the resulting data tends to approach a normal distribution. This indicates that the more scales used, the closer the data distribution pattern is to a normal distribution. Third, an increase in scale points is associated with a decrease in the resulting reliability coefficient. This is consistent with the findings by Wu and Leung (2017), which stated that adding scales to the Likert scale leads to a closer approximation of its underlying distribution.

Discussion

The results from Table 4 reveal several key trends: 1) Increasing the number of scale points correlates positively with increased data variance, indicating more significant variability with more scales. 2) With more scale points, the resulting data distribution tends to approximate a normal distribution more closely, suggesting a closer alignment to normality as scales increase. 3) However, a rise in scale points is linked to a decrease in the resulting reliability coefficient. This finding aligns with prior research by Wu and Leung (2017), which suggests that adding scales to the Likert scale brings it closer to its underlying distribution.

The primary objective of this study is to explore the impact of the number of scale points and score distribution on the reliability coefficient. Specifically, the study aims to enlighten judgment practitioners and evaluators about crucial aspects, determining the optimal number of scale points required to estimate

the reliability coefficient accurately. The variable of the number of scale points and the occurrence of non-normal distributions are identified as potential challenges in accurately estimating reliability coefficients. Previous studies focusing on the number of scale points and the effect of non-normal distributions on reliability coefficients have not thoroughly explored the impact of these variables on the reliability coefficients. Consequently, this study aims to comprehensively investigate the effect of each of these variables individually and in combination on the reliability coefficient.

Gaining a deeper understanding of the influence of the number of scale points and non-normal distributions on reliability coefficients can significantly aid practitioners in developing rubrics that yield more accurate estimates of measurement reliability. The outcomes of this study are intended to provide valuable insights for utilising performance assessment and holistic rubrics as decision-making tools for various contexts, such as high school final exams or licensure examinations based on written tests.

When analysing the outcomes, it is imperative to consider four key factors, especially in a sample size of 1000. Firstly, it is crucial to note that the reliability coefficient examined in this study specifically focuses on the number of scale points. Thus, the generalizability of the findings may not extend to situations involving multiple aspects. Secondly, it is essential to recognise that the study's results may not universally apply to the impact of non-normal distributions on reliability coefficients, excluding the normal distribution. Thirdly, the sample size of 200 participants, which reflects common practices in research studies and program evaluations, must be considered. Lastly, incorporating odd and even scales in the study was intentional, aiming to validate findings from previous studies.

One of the primary research questions addressed in this study focuses on examining the influence of the number of scale points on the reliability coefficient. Regarding the number of scale points, the findings of this study corroborate the conclusions drawn by prior researchers regarding the association between the number of scale points and the reliability coefficient. The majority of researchers have observed a positive relationship, wherein an increase in the number of scale points corresponds to an increase in the reliability coefficient (Arshad et al., 2022; Eisinga et al., 2012; Panayides, 2013; Raadt et al., 2021; Rahayu & Abidin, 2017; Shumate et al., 2007). In this study of the reliability coefficients, the best points seem to be around the 2- and 8-point scales, after which scaling up has little effect on the bias associated with the reliability coefficient estimates. This contradicts the results of research conducted by Postmes et al. (2013), which states that a scale with 7 points provides good reliability.

As alternative assessments such as portfolios and performance assignments have gained popularity in the past decade, the utilisation of a 4-to-6-point scale for grading these assessments has become increasingly prevalent. Consequently, researchers need to determine the suitability of the scores associated with such evaluations. Moreover, in many of these assessments, the distribution of scores may deviate from the normal distribution, potentially impacting the accuracy of the reliability measure. Educators and evaluators risk making decisions based on flawed measurements without clearly understanding the rubric's optimal number of scale points and the potential bias introduced by skewed score distributions. Thus, it is crucial to investigate these factors to ensure reliable outcomes.

The findings of this study reveal a notable pattern concerning the relationship between the number of scale points and the estimated reliability coefficient. With the increase in scale points, the reliability coefficient initially shows an upward trend, reaching its peak at a certain point. However, beyond this point, further increases in the number of scale points lead to a decrease in the reliability coefficient. Specifically, the estimated reliability coefficient for the 8-point scale is almost identical to the estimates using Cronbach's α, Guttman's λ2, and Guttman's λ6.

Furthermore, when comparing pairs of scale points, it is observed that the reliability coefficient for a 2-point scale surpasses that of a 3-point scale. Similarly, the reliability coefficient for a 5-point scale exceeds that of a 4-point scale, and the reliability coefficient for a 7-point scale is higher than that of a 6-point scale. Moreover, the reliability coefficient for an 8-point scale is greater than that of a 9-point scale, and the reliability coefficient for a 10-point scale surpasses that of an 11-point scale. Cronbach's α,

Guttman's λ2, and Guttman's λ6 all demonstrate that scales with an even number of points tend to yield higher reliability coefficient values than those with an odd number of points, as seen in Table 4.

The findings of this study have significant implications, as previous research has suggested that distinguishing between the meanings of different points on multiple-point scales can be challenging for most individuals. However, although the reliability coefficient based on the 8-point scale is higher than the coefficients based on the 10-point, 6-point, and 4-point scales, the difference in coefficient values is minimal, with a decrease of approximately 0.09. Consequently, practitioners evaluating assessments can obtain a nearly equivalent reliability estimate using a longer scale than an 8-point scale. Therefore, it can be concluded that including many scale points is not necessary to achieve the desired level of reliability according to reliability theory.

However, it is important to note that using a 4-to-6-point scale is commonly employed. Nevertheless, this study reveals that the reliability coefficient for a 5-point scale is slightly higher than that of the 4-point and 6-point scales. These findings align with the research conducted by Menold and Tausch (2015), which demonstrates that the rating scale's format influences the measurement quality. It is worth mentioning that Rouse's research (2015) suggests that the length of the scale does not impact the reliability of the scores. Additionally, other studies indicate that the level of reliability is only minimally influenced by the number of items (Piqueras et al., 2017).

The 11-point scale may be useful because it allows a higher level of precision but also places a high cognitive load on the respondent. This can result in a higher measurement error rate (Menold & Toepoel, 2022). The 11-point scale is useful because the analytical operations result in a more consistent assessment, i.e., higher reliability (De Beuckelaer et al., 2013). However, their results also show that a seven-point scale is a more reasonable alternative. A further reduction to a five-point scale is troublesome and results in a relatively high level of inconsistency in the answer scores. Altuna and Arslan (2016) show that reliability increases with larger scale points, although the increase is insignificant. These authors conclude that the 5-point scale may be easier to apply and preferable. Although the 5-point scale resulted in a higher non-response rate than the longer answer scale, the 11-point scale obtained a more positive evaluation of the questionnaire in the context of a mixed set (Toepoel & Funke, 2018). This finding carries considerable significance, as it is common for behavioural measures to deviate from a normal distribution pattern. It aligns with the results of previous studies, which showed that the estimated alpha coefficient was greatly increased in the presence of outliers, and like previous findings, the effect of outliers decreased with increasing theoretical reliability (Liu et al., 2010; Rahayu & Abidin, 2017; Shumate et al., 2007). Knowing that the distribution is skewed does not appear to overestimate or understate the value of the reliability coefficient, thereby reducing concerns over the estimation of the reliability of the assessment.

Understanding the influence of the number of scale points on the reliability of an instrument is crucial within the context of practical-based methods. This is because reliability measures the extent to which an instrument is consistent and accurate in measuring the intended construct. By employing practical-based methods, comprehension of the impact of the number of scale points on instrument reliability can aid in developing and selecting more effective instruments. In research, the appropriate selection of scale points can affect the quality of the collected data and the reliability of research findings.

In research, too few scale points can result in losing essential information and diminish the instrument's ability to capture variations within the measured construct. On the other hand, an excessive number of scale points can introduce unnecessary complexity, confuse respondents, and increase the likelihood of measurement errors. By understanding the influence of the number of scale points on instrument reliability, researchers or practitioners can make more informed and prudent decisions when selecting the appropriate number of scale points. In practical-based methods, selecting the correct scale points ensures that the resulting instrument provides reliable and accurate data for decision-making and practical actions.

Conclusion

The study concludes that while more scale points contribute to increased variance and a more normal distribution, it does not necessarily lead to higher reliability coefficients. Therefore, practitioners should consider the trade-off between the measurement precision and the scale length's practicality. The study highlights the importance of balancing the number of scale points to ensure reliable and accurate data for decision-making in various practical applications. The study provides valuable insights for educational decision-makers, emphasising the importance of selecting an optimal number of scale points for accurate and reliable assessments.

References

Alan, Ü., & Kabasakal, K. A. (2020). Effect of number of response options on the psychometric properties of Likert-type scales used with children. *Studies in Educational Evaluation*, *66*. https://doi.org/https://doi.org/10.1016/j.stueduc.2020.100895

Allen, M. J., & Yen, W. . (1979). *Introduction to measurement theory*. Brooks/Cole Publishing Company.

Altuna, O. K., & Arslan, F. M. (2016). Impact of the Number of Scale Points on Data Characteristics and Respondents' Evaluations: An Experimental Design Approach Using 5-Point and 7-Point Likert-type Scales. *İstanbul Üniversitesi Siyasal Bilgiler Fakültesi Dergisi*, *55*, 1–20. https://doi.org/10.17124/IUSIYASAL.320009

Arshad, S. S., Zaman, S., & Nazir, A. (2022). Development and Validation of Scale for Assessment of Followership among School Teachers. *International Journal of Instruction*, *15*(3), 1031–1046. https://doi.org/https://doi.org/10.29333/iji.2022.15355a

Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, *36*(1), 3–15. https://doi.org/10.1002/JOB.1960

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

De Beuckelaer, A., Toonen, S., & Davidov, E. (2013). On the optimal number of scale points in graded paired comparisons. *Quality & Quantity*, *47*(5), 2869–2882. https://doi.org/10.1007/S11135-012-9695-2

Ebel, R., & Frisbie, D. (1991). *Essential of Educational Measurement*. Prentice-H.

Eisinga, R., Grotenhuis, M., & Pelzer, B. (2012). The reliability of a two-item scale : Pearson, Cronbach or Spearman-Brown? *International Journal of Public Health*. https://doi.org/http://dx.doi.org/10.1007/s00038-012-0416-3

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-Analysis of Coefficient Alpha: A Reliability Generalisation Study. *Journal of Management Studies*, *55*(4), 583–618. https://doi.org/10.1111/JOMS.12328

Liu, Y., Wu, A. D., & Zumbo, B. D. (2010). The Impact of Outliers on Cronbach's Coefficient Alpha Estimate of Reliability: Ordinal/Rating Scale Item Responses. *Educational and Psychological Measurement*, *70*(1), 5–21. https://doi.org/10.1177/0013164409344548

McColly, W., & Remstad, R. (1965). Composition Rating Scales for General Merit: An Experimental Evaluation. *Journal of Educational Research*, *59*(2), 55–56. https://doi.org/10.1080/00220671.1965.10883300

Menold, N., & Tausch, A. (2015). Measurement of Latent Variables With Different Rating Scales: Testing Reliability and Measurement Equivalence by Varying the Verbalization and Number of Categories. *Http://Dx.Doi.Org/10.1177/0049124115583913*, *45*(4), 678–699. https://doi.org/10.1177/0049124115583913

Menold, N., & Toepoel, V. (2022). Do Different Devices Perform Equally Well with Different Numbers

of Scale Points and Response Formats? A test of measurement invariance and reliability: *Sociological Methods & Research*, 1–42. https://doi.org/10.1177/00491241221077237

Panayides, P. (2013). Coefficient Alpha: Interpret With Caution. *Europe's Journal of Psychology*, *9*(4), 687–696. https://doi.org/10.5964/ejop.v9i4.653

Piqueras, J. A., Martín-Vivar, M., Sandin, B., San Luis, C., & Pineda, D. (2017). The Revised Child Anxiety and Depression Scale: A systematic review and reliability generalisation meta-analysis. *Journal of Affective Disorders*, *218*, 153–169. https://doi.org/10.1016/J.JAD.2017.04.022

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology*, *52*(4), 597–617. https://doi.org/10.1111/BJSO.12006

Raadt, A. de, Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2021). A Comparison of Reliability Coefficients for Ordinal Rating Scales. *Rhode Island Medical Journal (2013)*, *38*, 519–543. https://doi.org/https://doi.org/10.1007/s00357-021-09386-5

Rahayu, W., & Abidin, Z. (2017). The Effect Number of Replication and the Number of Option Scale toward the Reliability Coefficient of Maximal in the Rubric Assessment of Vocational Learning Outcome. *American Journal of Education Research*, *5*(6), 645–649. https://doi.org/10.12691/education-5-6-9

Retnawati, H. (2020). *Validitas, Reliabilitas & Karakteristik Butir: Panduan untuk Peneliti, Mahasiswa, dan Psikometrian*. Parama Publishing.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304–307. https://doi.org/10.1016/J.CHB.2014.11.004

Shumate, S. R., Surles, J., Johnson, R. L., & Penny, J. (2007). The effects of the number of scale points and non-normality on the generalizability coefficient: A Monte Carlo study. *Applied Measurement in Education*, *20*(4), 357–376. https://doi.org/10.1080/08957340701429645

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/S11336-008-9101-0/TABLES/5

Toepoel, V., & Funke, F. (2018). Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies*, *25*(2), 112–122. https://doi.org/10.1080/08898480.2018.1439245

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*, 769. https://doi.org/10.3389/FPSYG.2016.00769/BIBTEX

Tsai, A. C., Liou, M., Simak, M., & Cheng, P. E. (2017). On hyperbolic transformations to normality. *Computational Statistics and Data Analysis*, *115*, 250–266. https://doi.org/10.1016/j.csda.2017.06.001

Wu, H., & Leung, S. O. (2017). Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *Journal of Social Service Research*, *43*(4), 527–532. https://doi.org/10.1080/01488376.2017.1329775

Wu, X. Z. (2020). Quantifying the non-normality of shear strength of geomaterials. *European Journal of Environmental and Civil Engineering*, *24*(6), 740–766. https://doi.org/10.1080/19648189.2017.1421102