JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 11(2), 2022, 166-180
DOI: http://dx.doi.org/10.15408/jp3i.v11i2.26591
http://journal.uinjkt.ac.id/index.php/jp3i

The Impact of *Z-Score* Transformation Scaling on the Validity, Reliability, and Measurement Error of Instrument SATS-36

Sumin[1], Fitri Sukmawati[2], Farida Agus Setiawati[3], Sumarin Asmawi[4]

Mathematics Tadris Study Program, Pontianak State Islamic Institute, Indonesia[1]
Islamic Psychology Study Program, Pontianak State Islamic Institute, Indonesia[2]
Doctoral Education Research and Evaluation Study Program, Yogyakarta State University, Indonesia[3]
Sultan Muhammad Syafiuddin Sambas Islamic Institute, Indonesia[4]

amien.ptk@gmail.com

Abstract

The Likert scale is a psychometric scale commonly used for response measurement. This measurement scale includes techniques for designing and administering surveys as well as coding and analyzing data. However, Likert scaling has various limitations that can affect the resulting data. This study aims to re-prove the number of dimensions of the SATS-36 instrument, prove the validity, and estimate the reliability of the statistical attitude instrument (SATS-36) on students at religious universities in Indonesia using *Z-Score* Transformation Scaling. The latent constructs of cognitive competence, value, difficulty, effect, and effort were constructed using a Likert scale according to the pattern of statements on each item. This study uses confirmatory research with a quantitative approach. For students at religious universities in Indonesia, 243 respondents were selected using a stratified one-stage cluster random sampling technique. Proof of validity and estimation of reliability was done using confirmatory factor analysis. The results of this study show that the rescaling method can improve the validity of the factors but cannot increase Cronbach's coefficient of internal consistency and cannot reduce the standard error of measurement for each item. This research implies that it is not enough to rescale or transform the data to improve the validity and reliability of a measuring instrument. However, it is necessary to calibrate the statement sentence or item question so that the item measures its construct. Further research also needs to test the effectiveness of rescaling in addition to the *Z-Score* in improving the validity and reliability of measuring instruments.

Keywords: reliability, rescaling, SATS-36, validity, *Z-Score*

Abstrak

Skala Likert *adalah skala psikometrik yang biasa digunakan untuk pengukuran respons. Skala pengukuran ini mencakup teknik untuk merancang dan mengelola survei serta pengkodean dan analisis data. Namun penskalaan Likert memiliki berbagai keterbatasan yang dapat mempengaruhi data yang dihasilkan. Penelitian ini bertujuan untuk membuktikan kembali jumlah dimensi instrumen SATS-36, membuktikan validitas, dan mengestimasi reliabilitas instrumen sikap statistik (SATS-36) pada mahasiswa perguruan tinggi keagamaan di Indonesia dengan menggunakan penskalaan transformasi* Z-Score*. Konstruk laten kompetensi kognitif, nilai, kesulitan, efek, dan usaha dikonstruksi menggunakan skala likert sesuai dengan pola pernyataan pada setiap item. Penelitian ini menggunakan penelitian konfirmatori dengan pendekatan kuantitatif. Untuk mahasiswa perguruan tinggi agama di Indonesia, dipilih 243 responden dengan menggunakan teknik pengambilan sampel acak klaster bertingkat satu tahap. Pembuktian validitas dan estimasi reliabilitas dilakukan dengan menggunakan analisis faktor konfirmatori. Hasil penelitian ini menunjukkan bahwa metode penskalaan dapat meningkatkan validitas faktor tetapi tidak dapat meningkatkan koefisien konsistensi internal* Cronbach Alpha *dan tidak dapat mengurangi kesalahan standar pengukuran untuk setiap item. Penelitian ini menyiratkan bahwa tidak cukup hanya mengubah skala atau mengubah data untuk meningkatkan validitas dan reliabilitas suatu alat ukur. Namun perlu dilakukan kalibrasi pada kalimat pernyataan atau butir soal agar butir soal tersebut mengukur konstruknya. Penelitian selanjutnya juga perlu menguji efektivitas penskalaan Z-Score dalam meningkatkan validitas dan reliabilitas alat ukur.*

Kata kunci: *reliabilitas,* rescaling*, SATS-36, validitas, Z-Score*

Introduction

Questionnaires are essential research instruments used in many studies. Questionnaires have been used in almost all fields of scientific study, as well as in the commercial and industrial sectors, because they offer many advantages, such as being more effective in data collection and more straightforward in collecting essential and sensitive information (Patten, 2016). Questionnaires and surveys may be excellent methods for gathering the information needed for research and assessment (Cox & Cox, 2008; Patten, 2016). To create a survey or questionnaire, the researcher must determine how the necessary data will be collected by selecting the appropriate scaling technique (Brace, 2018; Taherdoost, 2019). The Scaling technique is a field of measurement that involves the design and construction of measuring equipment in this sense (Clark & Watson, 2019).

The scaling technique is used to measure different psychological aspects like attitudes, perceptions, and preferences of people with the help of a predefined set of stimuli and instructions (Mehra, 2017). One of the most widely used scaling methods is the attitude scale for measuring instruments, and the Likert scale is applied as one of the most basic psychometric tools. It is often used in sociology, psychology, information systems, politics, economics, and other studies. In addition to offering flexibility in its use, the Likert scale also has many disadvantages. One of the significant drawbacks of the Likert technique is that a procedure for finding the neutral point has not been developed (Kandasamy et al., 2020; Pervez et al., 2020). Given the individual scores on the scale, it is not possible to determine whether the individual is "favorable" or "unfavorable" in addition, the use of closed response formats on the Likert scale forces respondents to make choices that may not match their answers, thereby potentially omitting or distorting information (Gillespie et al., 2021; Iwaniec, 2019).

Psychological tests usually include a response scale that aims to regulate and limit the choices available to the respondent and facilitate assessment. One such response scale is the Likert scale, introduced initially with a 5-response form, but in practice, it varies significantly in the nature and number of response options (Simms et al., 2019).

Although Rensis Likert (inventor of the Likert scale) assumes that the Likert scale has the quality of an interval scale, many experts consider the Likert scale to be ordinal because the interval scale requires that the differences between two successive scales reflect the same differences in the variables being measured. It is incorrect to assume that the intensity of “strongly disagree” and “disagree” is equal to the intensity of feelings between the other consecutive categories on a Likert scale (Li, 2013).

Contemporary social sciences and humanities education cannot avoid lectures related to research methodology and statistics. However, many students do not realize that they will be faced with many statistics courses when choosing social sciences and humanities. The gap between reality and expectations has caused students' anxiety about studying statistics in universities. Mismatches between expectations and curriculum can have both positive and negative effects, on the one hand, it can disrupt students' learning activities, but on the other hand, it can encourage them to study harder. The way students deal with this conflict depends on their level of statistical anxiety (Maloshonok & Terentev, 2017). This is in line with the study of Onwuegbuzie & Wilson (2003) that two-thirds to four-fifths of students feel anxious about statistics courses and research methods.

A study by Bose found that The Statistics Instructor is the mental image of a statistics teacher for the students. If a student feels their teacher conveys material that is difficult to understand, then the teacher will get a flawed assessment from students; otherwise, if a student feels the teacher is fun and the material is easy to understand during the course or meeting, the teacher gets an excellent perceptual assessment from the students (Bose, 2017), The study was supported by Saidi (2019) who investigated the validity

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

and reliability of the attitudes of rural junior high school students towards statistical learning, and the research uses an instrument developed by Schau (2003), and Ramirez et al. (2012) instrument items were constructed using a Likert scale, and the results of confirmatory factor analysis confirmed that six factors form student statistical attitudes consisting of 36 items resulting in 26 valid and reliable instrument items and ten other items invalid.

Based on the research results, several studies have measured students' attitudes towards learning materials, including learning science and statistics, such as studies conducted by Schau (2003), Osborne et al. (2003), Wang & Berlin, (2010), Schau et al.(2012), Lovelace & Brickman (2013), Pelch & McConnell (2017), Crouch et al. (2018), and others. According to a study conducted by Ramirez et al. (2012), there are at least 9 statistical attitude measurement models that have been developed, one of which is the STAT-36 instrument. However, this study focuses on re-verifying the number of instrument dimensions, proving factorial validity, construct validity, and measurement errors, and estimating the reliability of Schau's Survey of Attitudes Towards Statistics (2003) on students of Islamic Religious Universities in Indonesia by providing treatment in the form of a *Z-Score* transformation. For each accepted answer, a Likert scale was used. The selection of students from religious universities in Indonesia as participants was made in order to to reveal how well the perceptions of students with social studies backgrounds in learning statistics were, due to the previous study conducted by Saidi & Siew (2019), and the participants were selected from students majoring in science and technology who relatively preferred to study statistics.

The challenge often faced by researchers in psychometrics is creating a valid and reliable measuring instrument or scale. This is quite difficult for a beginner because it requires skills in compiling instrument statements that are genuinely able to measure the construct and choose the level of measurement needed because of the instrument item language and sentence form (positive or negative) style. Using the Likert scale affects the validity of the factors (Naji Qasem & Ahmad Gul, 2014). *Z-Score* statistics is a simple approach and easy to implement because it requires little statistical skill to convert or scale ordinal measurement levels to intervals or analytical skills to determine dimensions or prove validity and reliability.

One simple and quite popular method for rescaling is the *Z-Score*. According to Molugram et al. (2017), *Z-Score* gives units of equal distance between measurement and mean, random variable with normal distribution with mean 0 and standard deviation 1. *Z-Score* is the easiest method, even for beginners with low statistical skills, and it is very easy to use. These are the advantages of the *Z-Score* compared to other rescaling methods.

The research question is "Can the statistical attitude measurement instrument (SATS-36) items given the *Z-Score* transformation treatment produce a measurement instrument that is proven to be valid and reliable in measuring statistical attitudes of students at Religious Universities in Indonesia?" The selection of students from religious universities in Indonesia as participants was carried out with the reason to reveal how well the perceptions of students with social studies backgrounds in learning statistics were, due to the previous study conducted by Saidi & Siew (2019), and the participants were selected from students majoring in science and technology who relatively preferred to study statistics. In fact, students' attitudes towards statistics had not been explored in religious universities prior to this study. It is very important to highlight that socio-religious students have a good view of statistics in religious colleges. This is because so far there have been so many statistics courses for the social sciences and humanities which sometimes surprises students (Khavenson et al., 2012).

This study aims to prove the validity and estimate the reliability of the statistical attitude measurement instrument (SATS-36) for students at religious universities in Indonesia using the *Z-Score* rescaling technique.

Literature Review

Statistical Attitude Scale

A study conducted by Mehra (2017, p. 41) found that "scale can be defined as the process of measuring the quantitative aspects of a subjective or abstract concept." Two main types of scales used to measure respondents' attitudes are single-item and multi-item scales. The attitude scale measures attitudes towards individuals, objects, ideas, or objects. (Gure, 2015). According to Dwyer (1993) there are 4 types of attitude measurement scale, namely; Thurstone scale, Likert scale, Guttman scale, and semantic differential scale.

Why do some students naturally excel in math and statistics while others struggle with the same concepts? Why do people choose specific courses and career paths while pursuing a statistical background? The Student Attitudes Model to Statistics, created by Ramirez et al. (2012), is a comprehensive conceptual model that researchers may use to investigate these topics.

Measurement is fundamental to science, and the two most essential qualities associated with measurement are reliability and validity (Clark & Watson, 2019). Statistical attitude measurement developed by Schau (2003) adapted by Saidi & Siew (2019, p. 656) includes six factors; "cognitive competence, value, difficulty, influence, effort, and interest, with the following indicators (see Table 1).

Based on Table 1, as many as 36 items measuring students' attitudes towards learning statistics were constructed using a Likert scale. The results of the CFA analysis in Saidi & Siew's research (2019, p. 656) found that ten items had a loading factor of <0.7, namely, C1, V1, V2, V3, D4, D5, D6, D7, A2, and A5. Cronbach's alpha internal consistency reliability estimation shows a coefficient value above 0.7 on all factors, with an outstanding category. The results of the Saidi & Siew study (2019, p. 656) are similar to the previous study conducted by Vanhoof et al. (2011) using multidimensional confirmatory factor analysis with six factors, and there are 5 SAST-36 instrument items that have a loading factor value <0.5, namely items D3, D5, D6, D7 and V7 which produce an RMSEA model accuracy index of 0.059, NNFI = 0.94, CFI=0.95, BIC = 2150.4 and AIC = 1781.3.

According to Clark & Watson (2019) The primary purpose of developing a scale is to create a valid measure of the underlying construction. Therefore, to obtain maximum evidence of validity, it is necessary to pay attention to several things, including (a) clear conceptualization of target constructs, (b) overly inclusive initial set of items, (c) writing items in clear and precise words, (d) test item sets against closely related constructs, (e) choose the proper validation sample, (f) pay attention to unidimensionality over internal consistency.

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

Table 1. Factors and Indicators Forming the Statistical Attitude Construct (SATS-36).

Construct	Factors	Indicators	Items	Scale
Attitude towards learning statistics	*Cognitive Competence*	1. *Easy to understand statistical concepts.* 2. *Understand statistical formulas.* 3. *Being able to study statistics.* 4. *Not making many mathematical mistakes in learning statistics.* 5. *Having an idea of what is happening on the topic of statistics.* 6. *Difficulty understanding statistics*	C6 C5 C4 C3 C2 C1	Likert Scale
	Value	1. *Statistics are relevant in life.* 2. *Using statistics in everyday life.* 3. *Statistical conclusions are often presented in everyday life.* 4. *Having an application for statistical data analysis* 5. *Thinking statistics can be applied in life outside of work.* 6. *Valuable statistics for the typical professional.* 7. *Statistical skills will make the job easier.* 8. *Statistics are valuable.* 9. *Statistics are becoming a mandatory part of professional training.*	V9 V6 V7 V8 V5 V4 V3 V1 V2	Likert Scale
	Difficulty	1. *Statistical formulas are easy to understand.* 2. *Statistics is an easy subject.* 3. *Statistics is a subject that most people learn quickly.* 4. *Studying statistics does not require much discipline.* 5. *Statistics do not involve massive calculations.* 6. *Statistics are not very technical.* 7. *Most people don't need to learn a new way of thinking to do statistics.*	D1 D2 D3 D4 D5 D6 D7	Likert Scale
	Effect	1. *No stress in statistics class.* 2. *Liked statistics.* 3. *No frustration while taking statistical tests in class.* 4. *Not afraid of statistics.* 5. *Enjoy learning statistics topics.* 6. *Solve statistical problems safely.*	A4 A1 A3 A6 A5 A2	Likert Scale
	Effort	1. *Study hard in statistics topics.* 2. *Study hard for every statistical test.* 3. *Complete all statistical homework.* 4. *Attend every statistics class session.*	E2 E3 E1 E4	Likert Scale
	Interest	1. *Interest in understanding statistical information.* 2. *Interest in studying statistics.* 3. *Interest in using statistics.* 4. *Interest in being able to communicate statistical information to others.*	I3 I4 I2 I1	Likert Scale

Factors Affecting Validity and Reliability

Factors affecting the low estimation of internal consistency and validity are poorly written items with too large a measurement area, homogeneity of the test sample, time limits set in testing, item difficulty level, and instrument length (Lakshmi & Mohideen, 2013; Thanasegaran, 2009). In the context of the test instrument, Kinyua & Okunya (2014) argues that the factors that influence the validity and reliability of the test instrument prepared by the teacher are teacher experience, training in test construction and analysis, education level, use of Bloom's taxonomy, test moderation, and test duration affect the validity and reliability of the test.

Rescaling Z-Score

The *Z-Score* is a raw score that has been used differently based on its deviation from the mean value and is given in standard deviation units. *Z-Score* can also be referred to as a standard score. The standard score is presented as a *Z-Score*, with a new score distribution having a mean value of zero and a standard deviation of one. What is the point of having a score like this Z? When the number of items from one aspect to another is not the same, even though theoretically both aspects have the same weight, using the *Z-Score* can help in scoring. This is the case when using a *Z-Score* is useful (Azwar, 2016).

According to Walpole et al. (1993), and Molugaram et al. (2017) random variable with standard normal distribution Z can be defined by the equation:

$$Z = \frac{x_i - \mu}{\sigma} \quad (1)$$

where; x_i is the i-th data on a random variable X_i, μ is the normal random variable average X, and σ is the standard deviation of a normal variable. A random variable $Z = (x_i - \mu)/\sigma$ it can be said to spread according to the standard normal spread if its probability function is determined by:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}, (-\infty < Z < \infty) \quad (2)$$

The standard normally distributed random variable is denoted by $N(0,1)$. The standard normal distribution is also known as the distribution or normal distribution of units. Standardization of normal distribution makes it easier for us to know the area below the normal curve through the standard normal curve table, namely: $Z\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$ for various points along the X-axis. Area at between point Z_1, and Z_2 below the standard normal curve represents probability, where z lies between Z_1, and Z_2 denoted by $P(Z_1 \leq Z \leq Z_2)$, where the general equation of the normal curve is:

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}, (-\infty < Z < \infty) \quad (3)$$

If the corresponding total frequency is N, then

$$y = f(x) = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} \quad (4)$$

This equation is exactly the normal odds curve.

Methods

This study is confirmatory research with a quantitative approach (Jaeger & Halliday, 1998), we intend to prove the validity and estimate the reliability of the statistical attitude measurement tool for socio-religious students at Islamic religious colleges in Indonesia. A number of respondents, as many as 243, was selected using a stratified one-stage cluster random sampling technique. The variables of this research are students' attitudes towards statistics learning, which consists of 6 factors; [1] cognitive competence, [2] scores, [3] difficulty, [4] effect, [5] effort, and [6] interest), the measurement for factor 1 to factor 6 using a mixed Likert scale adjusted for statement patterns for each item, at the analysis stage, rescaling is carried out using the *Z-Score* transformation or better known as the summated rating. The research was conducted at Indonesia's Islamic religious universities in the social humanities (religious) studies field. We deliberately chose students from the social humanities field to avoid perception bias in science and technology students who generally like statistics or mathematics lessons. Data collection in this study used an online questionnaire via a google form. We analyzed the data to prove validity and reliability and estimated reliability using a structural equation modelling analysis package called *Lavaan* on the open-source R studio software. Before proving the validity and estimating reliability, we first re-proved the instrument dimensions using exploratory factor analysis software R studio (Brown, 2015; Harrington, 2009).

Results and Discussion

Results

Instrument Unidimensionality Checking

The dimensionality test of psychometric measuring instruments in this study was conducted using exploratory factor analysis (EFA) on the data of 125 respondents from a total sample of 243 respondents. EFA analysis includes sample adequacy test, the correlation between items in the same construct, assessment of factor loading, and visual dimension formation through scree plots. The assumption test of sample adequacy in exploratory factor analysis can be assessed using the Kaiser Meyer Olkin test (KMO test) statistic, provided that the value of Measure of Sampling Adequacy (MS) > 0.5, means; the assumption of the minimum number of samplings in the EFA analysis is met. The results of the KMO test, in this case, using the R Studio software are presented in Table 2 as follows:

Table 2. Sample Adequacy and Correlation Between Items.

Kaiser-Meyer-Olkin factor adequacy	
Overall MSA	0.87
Bartlet Test	
Chi-Square	3112.15
P Value	0,000
Df	630

Based on the MSA value, all instruments have an MSA value of 0.87 (MSA>0.5), and each item has an MSA value> 0.5, which means; The assumption of sample adequacy in the EFA analysis, in this case, is satisfied. The second assumption in the confirmatory factor analysis is the correlation between items. EFA analysis requires a significant correlation between measurement items so that these items can be grouped on certain factors. Testing the correlation between items using Bartlett's Test, if the probability of χ^2 is significant (P <0.05), it means that the items are significantly correlated.

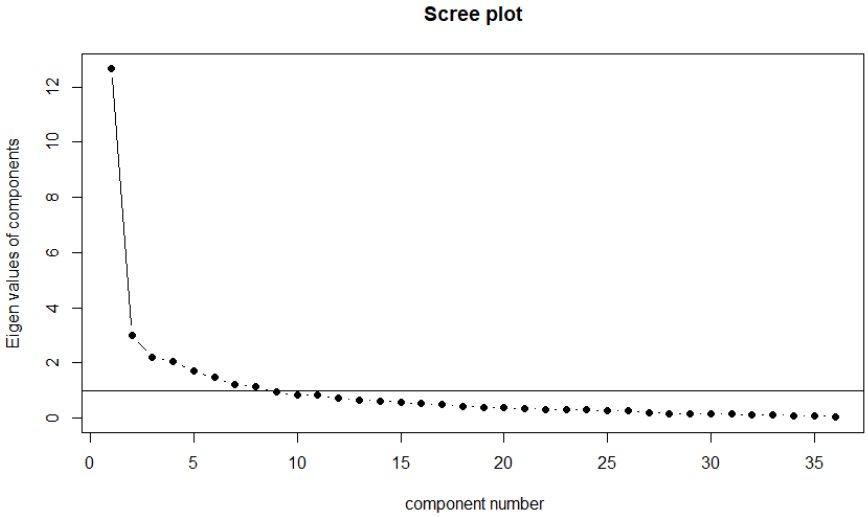

Figure 1. Scree Plot (Unidimensionality of Instruments).

Based on the scree plot image, inflation points occur at eigenvalue > 1 at 7 points, but the eigenvalue factor 1 is more than double factor 2, and so on, which can be seen visually. This shows that the dimensions or aspects formed have one dimension with seven factors. We can confirm the scree plot using the principal component analysis graph in Figure 2, which shows the eigenvalue of each item convergent towards dimension one as much as 35.25%, while dimension 2 is only 8.36%.

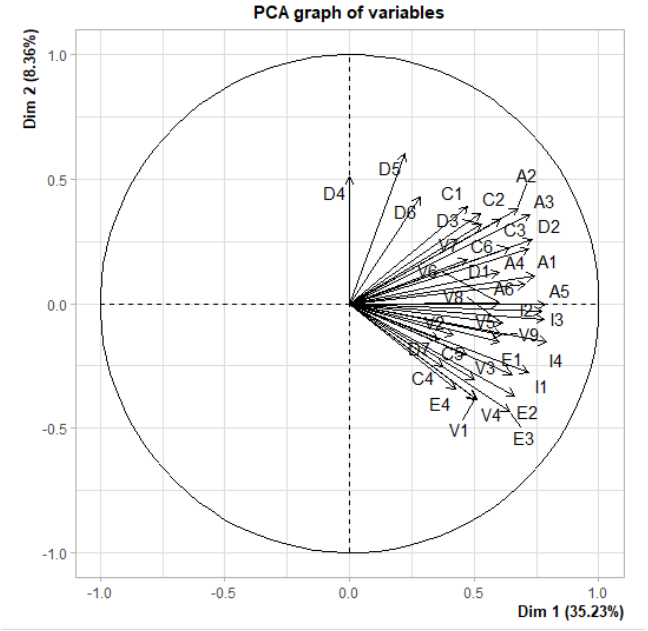

Figure 2. PCA Graphics (Unidimensionality)

Based on the results of the exploratory factor analysis above, verifying the scree plot in Figure 1, we can conclude that the SATS-36 instrument is a multifactor unidimensional instrument (consisting of 7 factors).

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

Proof of Validity

Through confirmatory factor analysis using the Study R software, we can prove the validity of the factors shown from 2 (two) measures, namely the size of the model accuracy (fit indices) and the loading factor.

Table 3. Fit Indices.

Criterion	Original	Z-Sore	Threshold	Remark
P-value (Chi-square)	0.000	0.000	>0.05	Poor fit
CFI	0.953	0.957	≥ 0.9	Good fit
RMSEA	0.054	0.052	≤ 0.08	Accepted
IFI	0.954	0.957	≥ 0.9	Good fit
GFI	0.898	0.900	≥ 0.9	Accepted
TLI	0.943	0.947	≥ 0.9	Accepted
PNFI	0.737	0.739	≥ 0.5	Good fit
SRMR	0.051	0.052	≤ 0.08	Accepted
AIC	7925.094	11365	Decreased	Accepted

Based on several fit index criteria in Table 3, on a Likert scale or original data or rescaling data using the *Z-Score* technique, the Chi-Square fit index shows a probability of <0.05, which indicates that the hypothetical model does not match the empirical models. However, RMSEA and SRMR showed values <0.08, TLI>0.9, and CFI>0.9. The hypothetical model is compatible with the empirical data because these three indices can be used to correct the Chi-Square weakness, which is very sensitive to the amount of data.

Overall, the accuracy index of the model on the response rescaled with *Z-Score* is better than the original data (Likert scale).

Table 4. Loading Factor.

	Original	***Z-Score***
Cognitive		
C1	0.731	0.729
C2	0.757	0.763
C3	0.722	0.712
C6	0.766	0.761
Value		
V1	0.751	0.725
V4	0.722	0.734
V5	0.71	0.699
V9	0.789	0.776
Difficulty		
D4	0.62	0.627
D5	0.753	0.744
D6	0.683	0.688
Effect		
A3	0.834	0.841
A4	0.874	0.874
A6	0.785	0.777
Effort		
E1	0.681	0.689
E2	0.904	0.913
E3	0.809	0.827
E4	***0.477***	***0.493***
Interest		
I1	0.714	0.709
I2	0.886	0.883
I3	0.919	0.913
I4	0.899	0.904

The value of the model accuracy index in Table 4 is strengthened by factor loading, which shows only E4 items that produce a factor loading of <0.5, meaning; in addition to item E4, other items used to measure students' statistical attitudes were significant in explaining latent cognitive constructs, value, difficulty, effort, and interest. Overall, the factor loading on rescaling data is higher than the factor loading before rescaling. However, several items show that the original data has a higher factor loading than after rescaling.

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

Table 5. Measurement Error Standards.

Item	Std. Measurement Error	
	Original	***Z-Score***
Cognitive		
C1	0	0
C2	0.098	0.096
C3	0.103	0.095
C6	0.097	0.096
Value		
V1	0	0
V4	0.081	***0.087***
V5	0.106	***0.117***
V9	0.106	***0.121***
Difficulty		
D4	0	0
D5	0.159	***0.161***
D6	0.121	***0.153***
Effect		
A3	0	0
A4	0.065	***0.066***
A6	0.067	***0.066***
Effort		
E1	0	0
E2	0.097	***0.113***
E3	0.092	***0.107***
E4	0.095	0.094
Interest		
I1	0	0
I2	0.086	0.086
I3	0.107	0.106
I4	0.117	0.109

The data collected after rescaling are shown to have a more significant standard error of measurement for each item than before the rescaling was performed, as shown in Table 5. However, numerous items reveal that the standard error of measurement in the original data (without rescaling) is higher than the rescaled data using the *Z-Score*. This is the case for seven of the 22 items utilized in this investigation.

Table 6. Estimated Reliability

	Size	Cognitive	Value	Difficulty	Effect	Effort	Interest
Original	alpha	0.832	0.844	0.725	0.868	0.795	0.916
Rescaling	alpha	0.830	0.835	0.730	0.868	0.817	0.914

Based on Table 6, the overall value of Cronbach's internal alpha consistency in the original data is higher than in the rescaling data. Based on the estimated reliability value, in this case, it shows that both the original data (Likert scale) and rescaling data have good internal consistency (alpha and omega > 0.7). However, the internal consistency of the original data is better than the data after rescaling using the *Z-Score*.

Discussion

Based on the results of proving the number of instrument dimensions using EFA, the findings of this study is slightly different from the results of the research by Saidi & Siew (2019), which found that ten items have loading factors below the threshold of 0.7, namely; C1, V1, V2, V3, D4, D5, D6, D7, A2, and A5, we use a minimum loading factor threshold of 0.5 and identify 14 items that have insignificant loading factors, namely; C4, C5, V1, V2, V3, V6, V7, V8, D1, D2, D3, A1, A2, and A5. The results of our study are also different from the findings of the Hommik & Luik (2017) study on the same instrument (STAT-36); Hommik & Luik (2017) found that there were 4 out of 27 items that had a factor loading of less than 0.5; they also investigated the effect of gender bias and education level on respondents' perceptions of statistics. Whereas our study did not investigate the effects of gender and education bias, future research may need to be strengthened by exploring the impact of gender and education bias.

The use of sentences in the questionnaire statement items affects differences in perceptions which are influenced by the level of education or understanding of each respondent. Differences in perception cause too high a diversity of responses between respondents, and this condition will have an impact on reducing the Z-Score transformation function; this is relevant to the theory of Kappal (2019) and Okunev (2022) which reveals that the Z-Score is designed for data that converges towards the distribution normal (no severe outliers). Another weakness is the Z-score statistical formula based on the calculated average. Standard deviation is part of parametric statistical techniques for interval or ratio data, so it is not suitable for use on Likert Scale questionnaire data that produces ordinal degrees of measurement. However, there is still debate about the degree of measurement on a Likert Scale.

Based on the results of the CFA analysis, the validity of the rescaling factor *Z-Score* succeeded in improving the fit indices but not significantly in increasing the loading factor. Rescaling *Z-Score* has also caused the standard error of the measurement to be higher than the original data. Likewise, with internal consistency, the rescaling method reduces the Cronbach's Alpha coefficient. The findings of this study strengthen the results of Khavenson et al. (2012) research which says that the STAT-36 instrument has items that are proven to be valid and reliable. However, this finding proves that the psychometric properties of Stat 36 items do not improve quality through the *Z-Score* rescaling process. *Z-Score* only transforms ordinal data into interval data, which is sometimes required in specific statistical analyses such as EFA analysis, CFA, or structural equation modelling. The findings of this study prove the study conducted by Didow Jr et al. (1985) who found that the scaling method could not improve the validity.

This study implies that in order to increase the validity and reliability of a measuring instrument, and it is not enough to scale back or transform the data, it is necessary to calibrate statement sentences or questions so that the items measure their constructs, expand the Likert scale response, for example from 5 to 7 or 9 which is in line with the results of research conducted by Malik et al. (2021). Researchers should not only rely on a quantitative approach to obtain evidence of high validity, reasonable reliability estimates, and low standard error of measurement but also be able to elaborate on and explore theories, concepts, constructs, and content suitability using qualitatively developed indicators. Further research also needs to test the effectiveness of rescaling in addition to the *Z-Score* in increasing the validity and reliability of measuring instruments.

Conclusion

Based on the analysis of EFA and CFA, we can conclude several essential things regarding the results of this study first; The results of this study are that the rescaling method can improve the validity of factors but cannot increase the coefficient of Cronbach's internal consistency and cannot reduce the standard error of measurement for each item. Based on this study, the use of *Z-Score* rescaling has no significant impact on improving the validity and reliability of the measuring instrument.

Acknowledgment

We want to thank all parties involved in this research and those who have contributed to the writing of this article, especially students from State Islamic Universities in Indonesia who have filled out the SATS-36 instrument, as well as those from the Doctoral Study Program of Education Research and Evaluation at the Yogyakarta State University, especially the supervisor of the Scaling Techniques course. We also thank everyone who assisted us in the writing of this article, whose names we cannot mention one by one here.

References

Azwar, S. (2016). Dasar-dasar psikometrika (II). *Pustaka Pelajar*.

Bose, D. (2017). *Development and validation of an instrument to measure attitude of undergraduate students towards Statistics*.

Brace, I. (2018). *Questionnaire design: How to plan, structure and write survey material for effective market research*. Kogan Page Publishers.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cox, J., & Cox, K. B. (2008). *Your opinion, please!: How to build the best questionnaires in the field of education*. Corwin Press.

Crouch, C. H., Wisittanawat, P., Cai, M., & Renninger, K. A. (2018). Life science students' attitudes, interest, and performance in introductory physics for life sciences: An exploratory study. *Physical Review Physics Education Research*, *14*(1), 10111.

Didow Jr, N. M., Keller, K. L., Barksdale Jr, H. C., & Franke, G. R. (1985). Improving measure quality by alternating least squares optimal scaling. *Journal of Marketing Research*, *22*(1), 30–40.

Dwyer, E. E. (1993). *Attitude scale construction: a review of the literature.*

Gillespie, B. J., Mulder, C. H., & Eggleston, C. M. (2021). Measuring migration motives with open-ended survey data: Methodological and conceptual issues. *Population, Space and Place*, *27*(6), e2448.

Gure, G. S. (2015). Different scale construction approaches used to attitude measurement in social science research. *International Journal of Research in Economics and Social Sciences*, *5*(1), 26–44.

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford university press.

Hommik, C., & Luik, P. (2017). Adapting the survey of attitudes towards statistics (SATS-36) for Estonian secondary school students. *Statistics Education Research Journal*, *16*(1), 228–239.

Iwaniec, J. (2019). Questionnaires: Implications for effective implementation. In *The Routledge handbook*

of research methods in applied linguistics (pp. 324–335). Routledge.

Jaeger, R. G., & Halliday, T. R. (1998). On confirmatory versus exploratory research. *Herpetologica*, S64–S66.

Kandasamy, I., Kandasamy, W. B., Obbineni, J. M., & Smarandache, F. (2020). Indeterminate Likert scale: feedback based on neutrosophy, its distance measures and clustering algorithm. *Soft Computing*, *24*(10), 7459–7468.

Khavenson, T., Orel, E., & Tryakshina, M. (2012). Adaptation of survey of attitudes towards statistics (SATS 36) for Russian sample. *Procedia-Social and Behavioral Sciences*, *46*, 2126–2129.

Kinyua, K., & Okunya, L. O. (2014). Validity and reliability of teacher-made tests: case study of year 11 physics in Nyahururu District of Kenya. *African Educational Research Journal*, *2*(2), 61–71.

Lakshmi, S., & Mohideen, M. A. (2013). Issues in reliabilityand validity of research. *International Journal of Management Research and Reviews*, *3*(4), 2752.

Lovelace, M., & Brickman, P. (2013). Best practices for measuring students' attitudes toward learning science. *CBE—Life Sciences Education*, *12*(4), 606–617.

Malik, M. A. A., Mustapha, M. F., Sobri, N. M., Abd Razak, N. F., Zaidi, M. N. M., Shukri, A. A., & Sham, M. A. L. Z. (2021). Optimal reliability and validity of measurement model in confirmatory factor analysis: different likert point scale experiment. *Journal of Contemporary Issues and Thought*, *11*, 105–112.

Maloshonok, N., & Terentev, E. (2017). The mismatch between student educational expectations and realities: prevalence, causes, and consequences. *European Journal of Higher Education*, *7*. https://doi.org/10.1080/21568235.2017.1348238

Mehra, S. (2017). Scaling techniques of attitude measurement. *International Journal of Advanced Education and Research*, *2*(2), 41–50.

Molugaram, K., Rao, G. S., Shah, A., & Davergave, N. (2017). *Statistical techniques for transportation engineering*. Butterworth-Heinemann.

Naji Qasem, M. A., & Ahmad Gul, S. B. (2014). Effect of items direction (positive or negative) on the factorial construction and criterion related validity in likert scale. *Khazar Journal of Humanities and Social Sciences*, *17*(3), 77–85. https://doi.org/10.5782/2223-2621.2014.17.3.77

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. *Teaching in Higher Education*, *8*(2), 195–209.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079.

Patten, M. (2016). *Questionnaire research: A practical guide*. Routledge.

Pelch, M. A., & McConnell, D. A. (2017). How does adding an emphasis on socioscientific issues influence student attitudes about science, its relevance, and their interpretations of sustainability? *Journal of Geoscience Education*, *65*(2), 203–214.

Pervez, A. K. M. K., Maniruzzaman, M., Shah, A. A., Nabi, N., & Ado, A. M. (2020). The meagerness of simple likert scale in assessing risk: how appropriate the fuzzy likert is? *NUST Journal of Social Sciences and Humanities*, *6*(2), 138–150.

Ramirez, C., Schau, C., & Emmioğlu, E. (2012). The importance of attitudes in statistics education. *Statistics Education Research Journal*, *11*(2), 57–71.

Saidi, S. S., & Siew, N. M. (2019). Investigating the validity and reliability of survey attitude towards statistics instrument among rural secondary school students. *International Journal of Educational Methodology*, *5*(4), 651–661.

Schau, C. (2003). Students' attitudes: The "other" important outcome in statistics education. *Proceedings of the Joint Statistical Meetings*, 3673–3681.

Schau, C., Millar, M., & Petocz, P. (2012). Research on attitudes towards statistics. *Statistics Education Research Journal*, *11*(2), 2–5.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Taherdoost, H. (2019). What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/Likert scale. *Hamed Taherdoost*, 1–10.

Thanasegaran, G. (2009). Reliability and validity issues in research. *Integration & Dissemination*, *4*.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). *Probability and statistics for engineers and scientists* (Vol. 5). Macmillan New York.

Wang, T., & Berlin, D. (2010). Construction and validation of an instrument to measure Taiwanese elementary students' attitudes toward their science class. *International Journal of Science Education*, *32*(18), 2413–2428.