Research Article https://doi.org/10.12973/ijem.8.4.625

International Journal of Educational Methodology
Volume 8, Issue 4, 625 - 637.
ISSN: 2469-9632
http://www.ijem.com/

Number of Response Options, Reliability, Validity, and Potential Bias in the Use of the Likert Scale Education and Social Science Research: A Literature Review

Imam Kusmaryono*
Universitas Islam Sultan Agung, INDONESIA

Dyana Wijayanti
Universitas Islam Sultan Agung, INDONESIA

Hevy Risqi Maharani
Universitas Islam Sultan Agung, INDONESIA

Received: May 29, 2022 ▪ Revised: July 14, 2022 ▪ Accepted: September 8, 2022

Abstract: This study reviews 60 papers using a Likert scale and published between 2012 – 2021. Screening for literature review uses the PRISMA method. The data analysis technique was carried out through data extraction, then synthesized in a structured manner using the narrative method. To achieve credible research results at the stage of the data collection and data analysis process, a group discussion forum (FGD) was conducted. The findings show that only 10% of studies use a measurement scale with an even answer choice category (4, 6, 8, or 10 choices). In general, (90%) of research uses a measurement instrument that involves a Likert scale with odd response choices (5, 7, 9, or 11) and the most popular researchers use a Likert scale with a total response of 5 points. The use of a rating scale with an odd number of responses of more than five points (especially on a seven-point scale) is the most effective in terms of reliability and validity coefficients, but if the researcher wants to direct respondents to one side, then a scale with an even number of responses (six points) is possible. more suitable. The presence of response bias and central tendency bias can affect the validity and reliability of the use of the Likert scale instrument.

Keywords: *Likert scale, literature review, potential bias, reliability and validity.*

To cite this article: Kusmaryono, I., Wijayanti, D., & Maharani, H. R. (2022). Number of response options, reliability, validity, and potential bias in the use of the likert scale education and social science research: A literature review. *International Journal of Educational Methodology*, *8*(4), 625-637. https://doi.org/10.12973/ijem.8.4.625

Introduction

A Likert scale is a form of scale used to collect data in order to find out or measure qualitative data (Boone & Boone, 2012; Cheng, 2012; Kokolakis, 2017). The data was obtained to determine a person's opinion, perception, or attitude towards a phenomenon (Kokolakis, 2017; Warmbrod, 2014). Currently, the Likert scale has been widely developed and used as a tool to conduct certain surveys, including in the field of education or social sciences where the data analyzed is more inclined to the form of quantitative data measurement (Bishop & Herron, 2015; Joshi et al., 2015).

Types of scales like Bogardus, Guttman, Likert, and Thurstone scales and others that we need to understand well. This is so that we have no difficulty when measuring the research data using this scale. However, the Likert scale is the most popular type of measurement scale and is widely used to measure attitudes (affective) in educational research or social science (Bishop & Herron, 2015).

On the other hand, there are still differences of opinion about the Likert scale which causes confusion for novice researchers (Guerra et al., 2016; Joshi et al., 2015; Subedi, 2016; Taherdoost, 2019). Is the data collected through the Likert scale in the form of ordinal or interval data types? Therefore, we need to understand the form of the Likert scale (including the calculation and the purpose of data collection) before deciding to use the Likert scale in the preparation of instruments or analysis of research data (Sullivan & Artino, 2013). The effectiveness of the Likert scale is also strongly influenced by the form of the question items in the questionnaire and the data analysis technique of the Likert scale (Joshi et al., 2015; Nemoto & Beglar, 2014).

The use of the Likert scale in research has involved many fields of research and has been published openly. The availability of abundant articles in online journals as reference material requires the intelligence of researchers in

* **Corresponding author:**
Imam Kusmaryono, Universitas Islam Sultan Agung, Semarang, Indonesia. ✉ kusmaryono@unissula.ac.id

© 2022 The Author(s). **Open Access** - This article is under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

referring to sources of articles that support the results of their research. The use of a Likert scale as an instrument or inappropriate data analysis will cause bias (Pimentel, 2019). The emergence of bias in data analysis resulted in the research results not being in accordance with the actual reality. Therefore, we feel the need to review several articles that use the Likert scale in research, especially research in the fields of education and social science.

In 2014 Robert Warmbrod conducted a literature review on the Likert scale. He reviewed 344 articles related to the Likert scale published in international journals between 1995 – 2012 (Warmbrod, 2014). His literature review focuses on the interpretation of scores on a Likert type scale. Considering current research developments and the lack of information about the 2012-2022 Likert scale literature studies, it is necessary to have up-to-date information on the use of Likert scales in education and social science research.

Literature Review

Several terms in this terminology are used to provide a common understanding of the focus of Likert scale research. The purpose of affirming the term is so that there is no shift in meaning from the original Likert scale. Terms, words, and word combinations in the context of this paper are presented in Table 1.

Table 1. Likert Scale Terminology (Lionello et al., 2021; Warmbrod, 2014).

Term		Description
Likert scale	:	The rating scale is relative to one perception of a phenomenon (object) given by respondents in stages with categories ranging from "strongly disagree" on one pole to "strongly agree" on the other pole.
Likert scale items	:	A single question that uses several aspects of the response alternatives. It is a fact that the Likert scale is always composed of several Likert items.
Likert response choice category	:	The label applied is in the form of response choice categories (e.g., "strongly disagree", "disagree", "neutral", disagree", "agree", "strongly agree").
Likert scale metric	:	A geometric function that determines the range of the same perception interval between its points on a Likert scale, constructed along with different directions on a continuum of space.
Likert Rate	:	The numerical value assigned to each category of Likert response options (e.g., 1-5 if considered as equidistant intervals).

Initially, the Likert scale developed by Rensis Likert (1932) used response choice categories 3 and 5. Response choices 3 were: agree, undecided (neutral), disagree. Response options 5 are: strongly agree, agree, neutral, disagree, and strongly disagree (Likert, 1932). The development of research in the fields of social science, education, and psychology influenced the use of the Likert scale as data analysis.

The Likert scale assumes that the intensity (strength) of an attitude is always linear i.e., on a continuum from strongly agree to strongly disagree, and makes the assumption that attitudes can be measured (Likert, 1932). Likert scales with odd response choice categories (5, 7, 9, and 11) are generally concentrated in the middle of the scale and lead to more items whose weights are assigned mostly in the middle of verbal descriptions.

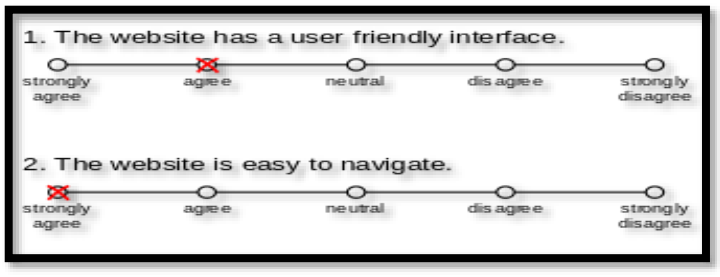

Figure 1. Likert Scale Items with Five Response Options

In several studies, other researchers used a Likert scale with the category of response choices even (4, 6, 8, or 10) (Jeong et al., 2019; Nemoto & Beglar, 2014; Taherdoost, 2019). This choice category excludes the middle (neutral) option on the grounds that they can force the respondent to give a positive response or a negative response so that it is easy to interpret (Baka et al., 2012). The characteristics of using a Likert scale with an even response choice category will show a very small difference in the mean in terms of variation (DeCastellarnau, 2018; Pimentel, 2019).

Item	Questions	Response			
		Strongly Disagree	Disagree	Agree	Strongly Agree
1	Teacher service in the classroom is open to all students				
2	Teachers help solve learning difficulties faced by students				

Figure 2. Likert Scale Matrix with Four Response Options

Figures 1 and 2 show the differences in intervals and verbal descriptions in the preparation of the number of answer choice categories. The figure (table) can be used as a guide or a tool to illustrate verbal descriptions of weighted averages calculated for percentages (considered valid) and have less prevention of errors (bias) that often occur in the practice of social science research and education.

Why does it matter if the Likert scale is an interval or ordinal type of scale? This question becomes an urgent matter to understand when the researcher intends to calculate the average score and perform certain statistical analyzes on the data collected from the Likert scale. The measurement scale is a set of rules for 'quantitating' data from the measurement of a variable (Çıplak & Çam, 2019; DeCastellarnau, 2018; Pimentel, 2019). In performing statistical analysis, the different types of data greatly affect the selection of models or statistical test tools. Not arbitrary data types can be used by certain test equipment. Furthermore, the statistical test depends on whether the Likert scale is an ordinal or an interval scale.

Pay attention to the Likert scale items in Figure 1 with five response options. Likert scale items on positive statements, then the response strongly agrees is definitely higher than the response agrees; the agreeable response is definitely higher than the neutral response; the neutral response is definitely higher than the disagree response; while the response to disagree is definitely higher than the response to strongly disagree. However, the response distance between strongly agree to agree and between agreeing to neutral and so on is certainly not the same and is not known with certainty. Therefore, the data generated by the Likert scale is ordinal data (Boone & Boone, 2012; DeCastellarnau, 2018).

While the scoring method where there is a weighting (Likert scale value): strongly agree = 5, agree = 4, neutral = 3, disagree = 2, and strongly disagree = 1 is just code to know which one is higher and which one is lower. The scoring method cannot be interpreted that strongly agree (5) being the same as neutral (3) plus disagree (2). This is in accordance with the characteristics of ordinal data, that ordinal data cannot be subjected to mathematical operations (DeCastellarnau, 2018; Mishra et al., 2018), but many researchers at the time of scoring from the Likert scale add up the scores for each item even though it is clear that the ordinal data scale cannot be added up.

Data with ordinal type is only to describe the summary of data as the frequency or percentage of responses in each category of Likert items. We can only use the data as the median or mode, and not the mean as a measure of central tendency (Jamieson, 2004). If so, in order for this data to be used in further analysis, the Likert data must be converted to interval data. The method that is often used is the method of successive interval (MSI) (Mondiana et al., 2018). A method that considers qualitative values as quantitative data in order to accept the statistical analysis. Numerical scores were assigned to each Likert item with potential choices and the average for all responses was calculated at the end of the survey. Then this Likert scale will change to an interval data type after going through the MSI (Mondiana et al., 2018; Solimun et al., 2017).

In the field of psychology, the Thurstone and Guttman scale used to be intervals (Aini et al., 2018). Then Rensis Likert developed the Likert scale into a scale with a response choice of 7,9,11 and so on as long as it is odd and there is neutral (Guerra et al., 2016; Taherdoost, 2019). Likert conducted research and the Likert questionnaire was changed in the form of a Thurstone and Guttman scale and then asked the same respondent, it turned out that the correlation value between the Likert scale and Guttman and Thurstone scales was 0.92 (Likert, 1932). So with the results of this study, the Likert scale can be considered a type of interval scale. Based on this reason, it is not surprising that research results published in well-known international journals (journal of marketing, journal of consumer behavior, journal of physiology, journal of human resources) do not transform data because they already view the Likert scale as an interval scale.

The Likert scale is the most widely used psychometric scale in survey research. The name of this scale is taken from a psychologist named Rensis Likert (1932), who published a report describing attitude measurement techniques and instruments for measuring constructs that describe psychological and social phenomena.

Likert scale is a bipolar scale to measure positive or negative responses to a statement (object) (DeCastellarnau, 2018; Kyriazos & Stalikas, 2018). The Likert scale type consists of statement items to define the content and meaning of the

construct being measured (Joshi et al., 2015; Sullivan & Artino, 2013). The response continuum to each statement is a tiered linear scale to indicate the extent to which respondents agree or disagree with certain social phenomena. To determine the quantification of the constructs of each individual, it can be calculated by adding up the individual response scores for each statement item. Statements that do not support the construct (negative statement), when measuring the construct, the response choice scores are reversed then the scores of several items on the scale are added up (Warmbrod, 2014).

Education researchers often use a Likert scale to measure the attitude of a person or group of people about perceptions of educational program policies. For example, interests, benefits, barriers, and challenges to learning practices; teacher performance; performance and service satisfaction; and self-perceptions about the level of student competence. Table 2 shows examples of constructs in the preparation of a questionnaire involving a Likert scale in educational research.

Table 2. Example of Constructs Measured on a Likert Scale in Educational Research

Construct	:	Perceptions of distance learning (20 items)
Response continuum	:	1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree
Target respondents	:	Elementary and middle school teachers

This review of the literature on the Likert scale aims to (a) critically review the concerns about the use of the Likert scale in data analysis of education and social science research, (b) critically review the reliability of the Likert scale in each category of even and odd response options, and (c) provide practical and appropriate solutions to reduce bias in the analysis of research data involving a Likert scale. This literature study is valuable and up-to-date supplementary information for researchers working on a Likert scale through a literature review between 2012 - 2022.

Methodology

Research Design

This study is a systematic review of the literature (Khalaf & Zin, 2018; Martins & Gorschek, 2016) which aims to analyze the use of the Likert scale in educational and social science research articles. Systematic Review is a method that uses a review, analysis, structured evaluation, classification, and categorization of previously produced evidence-based evidence (Ahn & Kang, 2018). The systematic review process is strictly limited to inclusion criteria (Martins & Gorschek, 2016).

Inclusion Criteria

Inclusion criteria are general characteristics of research subjects from a target population that is stretched and will be studied (Ahn & Kang, 2018). The inclusion criteria determined from this study are: (a) research articles involving the use of the Likert scale, (b) research with 20 or more respondents, and (c) research articles published in educational or social science journals between 2012 to 2021.

Tracking Technique and Screening

Literature related to research data according to inclusion criteria was tracked online from the database of indexed journals Scopus, ERIC, and other websites. The keywords used to track were attitude scale, Likert scale, and affective assessment. The digests taken as research data (articles) are the title of the study, name of the researcher, year of publication, place of research, sample size, research methods, and research results with significant values. Screening to review this literature, we used the preferred reporting items for the systematic review and meta-analysis (PRISMA) method (Selcuk, 2019; Warmbrod, 2014). The PRISMA flowchart and the literature review process are shown in Figure 3.

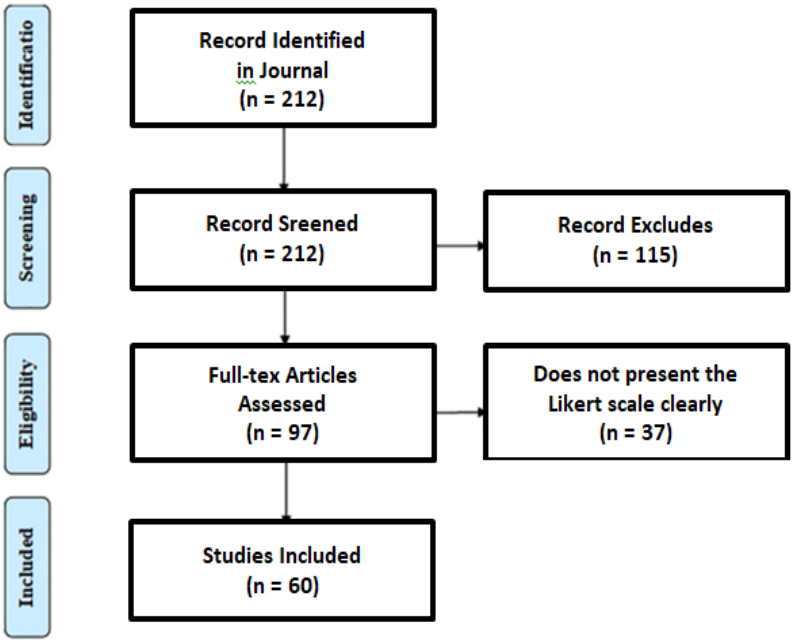

Figure 3. PRISMA Flowchart of Review Process

Figure 3 shows that based on the inclusion criteria starting with the work of (a) tracking related papers online (2012 – 2021) related to the scale as in the research report (n = 212), (b) filtering papers by Likert scale keywords, (c) search results specifically exclude papers (n = 115), (d) determine which papers have priority (n = 97) to be studied in-depth and then discuss clearly the Likert scale specifically exclude (n = 37), and (e) determine the papers that fall into the category to be studied (n = 60).

Data Analysis

The data analysis technique in this systematic review research is data extraction. Data extraction is done by taking all research data obtained from scientific journals used for research. Then, the researcher changed the data obtained into new data by filtering the data into several categories (Mathes et al., 2017; Munn et al., 2014; Pedder et al., 2016; Schmidt et al., 2021). Researchers only take valid data and do not include data that is less valid so that optimal new data and satisfactory results are obtained. Data extraction is the most important phase in research using a systematic literature review method (Jonnalagadda et al., 2015)**.** This phase is very vulnerable to a lot of research data that may be lost, if not careful in filtering the data (Mathes et al., 2017).

All data taken from journal searches were extracted according to the research objectives. The main data taken from the journal article include researchers and research year, research design, research location, number and characteristics of research samples, questionnaire instrument with Likert scale, and research results and conclusions. The data is entered in the data extraction form and displayed in the form of a table (Mathes et al., 2017; Popenoe et al., 2021).

Data Synthesis

In this systematic literature review research, data (including articles) were synthesized in a structured manner using the narrative method (Mathes et al., 2017). Narrative synthesis is done by grouping the extracted data (similar) according to the measured results to answer the research objectives (Thomas & Harden, 2008). The data that has been collected is then looked for similarities and differences are discussed to draw conclusions (Munn et al., 2014; Onwuegbuzie et al., 2012).

Credibility

Researchers have collected real data in the field and interpreted the authentic data accurately to provide assurance that reliable research has credible attributes (internal validity) (Miles et al., 2014). In order to achieve credible research at the stage of the data collection and data analysis process, a forum group discussion (FGD) was conducted. FGD members consist of researchers and experts in the field of psychology. FGD aims to get input and suggestions in data analysis. In addition, to increase credibility, the results of the analysis are compared with the findings of previous experts.

Findings / Results

Tracking the data sources obtained as many as 60 article titles have been obtained that match the specified inclusion criteria. The use of the Likert scale in this study is manifested in the form of a survey in which a questionnaire is given to respondents. The results of tracking manuscripts based on our source of variance are presented in Table 3.

Table 3. Tracking Results Based on Research Variant Sources

Source of Variance	Article Publication (n)	Distribution Technique Questionnaire	Number of Response Options
Social science journal	28	---	---
Educational journal	32	---	---
Paper and pencil method	---	18	---
Online form method	---	42	---
Even option Likert scale	---	---	6
Odd option Likert scale	---	---	54
Total	60	60	60

Based on Table 3, it can be explained that the need for using a Likert scale in social science research (47%) and educational research (53%) is almost balanced. The system for distributing questionnaires to respondents is known (30%) conventionally distributing questionnaires through the paper and pencil method. Most (70%) were done online via e-mail, telephone, WhatsApp, Google form (Google drive), and other online media.

The results of careful checking (Table 3) were obtained that only 10% of studies used a measurement scale with an even response choice category (4, 6, 8, or 10 choices) (Dawes, 2018; Guerra et al., 2016; Krosnick & Holbrook, 2012; Mircioiu & Atkinson, 2017; Nemoto & Beglar, 2014**;** Pimentel, 2019). But in general (90%) of research uses a measurement instrument that involves a Likert scale with odd response choices. Apparently, the Likert scale with a response choice of 5 (five) (e.g., Alrajeh & Shindel, 2020; Dilekli & Tezci, 2019; Hartley, 2013; Martín et al., 2018; Taherdoost, 2019; Ulia & Kusmaryono, 2021; et al.) is still the most popular than the other odd (7, 9, or 11) response choices (James, 2019; Lewis & Erdinç, 2017; Martín et al., 2018; Sirganci & Uyumaz, 2021)**.**

This literature review research is focused on the variables (a) the number of response points on the Likert item (Likert scale), (b) reliability and validity of the survey instrument with a Likert scale, and (c) the potential for bias in survey research. The results of investigations on research variables from journal articles that are included in the literature review research category are presented in Table 4.

Table 4. Investigation of Research Variables

Research Variable	Number of Participant	Number of Question Items	Data Sources
Even response point (Likert-type-scale: 4, 6, 8, or 10 point)	30 until 536	30 until 60	(Dawes, 2018; Guerra et al., 2016) (Krosnick & Holbrook, 2012) (Mircioiu & Atkinson, 2017) (Nemoto & Beglar, 2014)
Odd response point (Likert-type-scale: 5)	30 until 7261	30 until 60	(Alrajeh & Shindel, 2020; Dilekli & Tezci, 2019; Hartley, 2013; Martín et al., 2018; Taherdoost, 2019; Ulia & Kusmaryono, 2021) (Boone & Boone, 2012) (Carey et al., 2017) (Józsa & Morgan, 2017) (Pimentel, 2019) (Cheng, 2012)
Odd response point (Likert-type-scala: 7, 9, 11, or 101)	332 until 1000	30 until 60	(James, 2019) (Lewis & Erdinç, 2017) (Martín et al., 2018) (Sirganci & Uyumaz, 2021)
Validity and reliability	30 until 473	32 until 50	(Bidermana & Reddockb, 2012) (Bolarinwa, 2015) (Carey et al., 2017) (Çetin et al., 2020) (Ciplak & Cam, 2019) (Cheng, 2012) (Chen & Liu, 2020) (Krosnick & Holbrook, 2012) (Sangwan et al., 2021) (Simms et al., 2019) (Taherdoost, 2016) (Zhang et al., 2021)
Bias factor	30 until 609	24 until 60	(Acosta et al., 2020) (Kreitchmann et al., 2019) (Krosnick & Holbrook, 2012) (Zumsteg et al., 2012) (Pimentel, 2019) (Xiong et al., 2020)

Table 4 shows that in general practice the use of a Likert scale with an odd number of response categories (Likert-type-scale 5, 7, and 9) is preferred by researchers in the social sciences (e.g., Alrajeh & Shindel, 2020; Dilekli & Tezci, 2019;

Hartley, 2013; Martín et al., 2018; Taherdoost, 2019; Ulia & Kusmaryono, 2021; et al.) They are attracted by the odd response in the middle (neutral or no opinion). In a recent empirical study, the use of a Likert scale with the number of response categories 5 or 7 resulted in a higher mean score for the maximum possible score and the difference was significant (Kyriazos & Stalikas, 2018; Simms et al., 2019). Meanwhile, psychometrics prefer a Likert scale with a category of 7 or 9. The reason is that they can stretch the intervals in order to get a more detailed and clear view (preference) of respondents.

Other researchers prefer to use a Likert force-number scale, namely the response category with even choices (4, 6, 8, or 10) (Dawes, 2018; Guerra et al., 2016; Krosnick & Holbrook, 2012; Mircioiu & Atkinson, 2017; Nemoto & Beglar, 2014; Pimentel, 2019). They reasoned to avoid the middle response (neutral/did not decide) as in the 5 responses choice Likert scale. Characteristics of the use of the even response choice Liker scale show very small differences between format scales in terms of variations in mean, kurtosis, or skewness (Pimentel, 2019). There are findings that some researchers do not test (do not present) the reliability and validity of the research instruments used to obtain data from respondents (Guerra et al., 2016; Pimentel, 2019; Sullivan & Artino, 2013).

Discussion

Likert Scale with Response Choice Even or Odd?

An important issue for researchers to pay attention to in developing an attitude rating scale is the number of answer choices on the questionnaire. Which one is suitable to be developed, a scale with an even or odd number of responses (Nadler et al., 2015). Based on an attitude scale reference with an even response choice category (4, 6, 8, or 10) (Dawes, 2018; Nemoto & Beglar, 2014) produces ordinal data types (Kyriazos & Stalikas, 2018). This assumption is caused on a scale with an even number of responses there is no "neutral" middle response option. Researchers did not collect gradations of negative or positive responses that could be obtained with a minimum of four choices (DeCastellarnau, 2018). If the answer choice does not have a midpoint even though it is labeled with a number then this only functions as an attribute that cannot be operated on mathematically.

Experts who look at Likert scales that have more than five response options allow the gradation to move more smoothly than negative or positive responses (James, 2019; Lewis & Erdinç, 2017; Taherdoost, 2016). This makes sense because an increase in the number of response choices on a Likert scale will lead to an increase in data quality (DeCastellarnau, 2018; Simms, et al., 2019; Taherdoost, 2019). According to Preston and Colman (2000) in general, respondents (participants) prefer a scale with an odd number of responses rather than an even number of responses. This is defined as the midpoint as a neutral point that will prevent respondents from being forced to take sides (Preston & Colman, 2000).

In addition to the five response choices in its development the Likert scale can also be used as a scale with seven or nine, or eleven answer choices (Joshi et al., 2015; Taherdoost, 2019). An empirical study found that some statistical characteristics of questionnaire results with various odd answer choices were very similar (DeCastellarnau, 2018; Simms et al., 2019; Taherdoost, 2019). Basically, the number of response categories will affect the psychological distance between categories, especially the most striking on the 7-point response scale. On a response scale of 7, it can be seen that when the number of categories is increased, the end point tends to approach the midpoint of the scale (Sirganci & Uyumaz, 2021).

The difference between a Likert scale with even and odd response options is whether a neutral response option as the midpoint is included or not (Martín et al., 2018). The existence of a neutral response choice as the midpoint becomes a significant difference in treating the Likert scale as an interval scale if the survey respondents use the midpoint of the scale for the true meaning of neutral. The assumption of interval scale influences the researcher's decision for descriptive statistical analysis (mean, median, mode, standard deviation, frequency, percentage, and inferential statistical analysis (Chen & Liu, 2020). On this occasion, we recommend the use of a rating scale with an odd number of responses of seven points. However, if the researcher wants to direct the respondent to one side, then a scale with an even number of responses of six points may be more suitable.

Potential Bias

The Likert scale with an odd number of response choice categories (neutral middle choice) and even response choices, both have the potential to cause serious bias problems if researchers are not prepared to anticipate their impact. If the Likert scale with odd response choices has a middle (neutral) (Ahn & Kang, 2018; Pimentel, 2019), then there is a possibility that respondents avoid extreme choices (strongly agree or strongly disagree) and choose to take a "safe" attitude, which is neutral. For example, the researcher wants to measure learning interest or learning motivation from a group of students, where the average questionnaire results show a "neutral" or "indecisive" attitude, the results of the questionnaire will be difficult to interpret. Although for respondents, the choice of “neutral” could mean 'don't care; and for others, it may state 'no opinion' (Krosnick & Holbrook, 2012). Cases like this will still confuse researchers in compiling research conclusions. When the respondent consciously does this by choosing the option in the middle of the scale, namely “neutral” or “safe”, a central tendency bias phenomenon occurs (Malone et al., 2014). If the researcher

knows that the majority of respondents will give a "disagree" or "neutral" response just for the purpose of avoiding a Likert item, then the "neutral" option should be omitted. Eliminating the “neutral” option will not compromise the reliability of the given answer (Krosnick & Holbrook, 2012).

Likert with an even number of answer choice categories has the potential to guide respondents to choose extreme answers (strongly agree or strongly disagree). Likert scale with an even response choice (without a neutral option) tends to force respondents to choose an opinion that is unequivocally "agree" or "disagree" (Moors et al., 2014). Moreover, if this attitude scale questionnaire is designed by someone who has a strong influence on the respondent, then the respondent tends to give answers that are not in accordance with their circumstances. Usually, respondents give extreme answers, namely "strongly agree" for positive questions or "strongly disagree" for negative questions. This phenomenon occurs where the tendency of individuals to respond to statement items by basing themselves on what is considered appropriate by society (social desirability), is referred to as respondent bias (Malone et al., 2014; Moors et al., 2014; Xiong et al., 2020; Zumsteg et al., 2012).

The tendency of respondents to provide answers to certain choices can be analyzed through the application of item response theory (IRT) (Dogan, 2018; Jeong et al., 2020; Thorpe & Favia, 2016; Tijmstra et al., 2018; Zanon et al., 2016). The results of this IRT analysis produce ordinal data instead of interval data. Analysis with this model allows item analysis with a deeper examination of precision across the score continuum. An illustration in the form of a graph of the results of attitude measurement with the application of IRT can be seen in Figure 4.

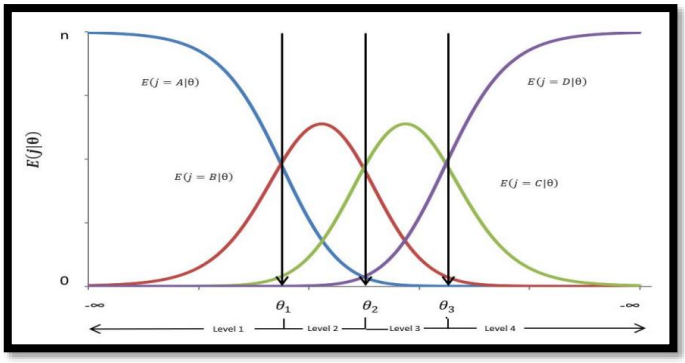

Figure 4. Item Characteristic Curve for a Scale Consisting of n Items with Four Answers

The item characteristic curve (ICC) in Figure 4 is a logistic function that shows the probability and item characteristics (on a Likert scale) of respondents supporting a particular response choice (to the extreme). For a scale with four response options, it is calibrated through a partial credit model where the y-axis is a function $P(X_i = j \mid \theta)$ which shows the probability that the respondent will choose the j-th answer to answer the i-item choice.

Look at Figure 4, after the data is plotted for each item (according to the respondent's answers) the resulting scale characteristic curve (SCC) is generated where there are four curves (four response options) that intersect at three points. The three points of intersection divide the scale into four levels. Level 1 = very negative, level 2 = negative, level 3 = positive, and level 4 = very positive. Based on the SCC it can be seen that there is a response bias in the choice of extreme response (Dogan, 2018; Thorpe & Favia, 2016).

Some researchers use a random intercept model based on the item response theory (IRT) model in an effort to control response bias that is prone to response styles (bias) such as social desirability (SDR) and acquiescent response (ACQ) (Jeong et al., 2020; Kreitchmann et al., 2019; Tijmstra et al., 2018; Zanon et al., 2016). By Type, SDR refers to the tendency (respondents) to respond in a way that is consistent with what others want. Type ACQ refers to a preference for responding positively on the rating scale, regardless of the content of the survey item. Conceptually related to ACQ, the opposite trend is usually called dis-acquiescence (Pimentel, 2019). Therefore, correcting and minimizing response bias can be done by exploratory methods (Ferrando et al., 2009). More specifically, response bias and central tendency bias can affect the variability and reliability of the use of the instrument Likert scale (Lozano et al., 2008; Moors et al., 2014).

Reliability

Reliability tests on psychometric scales generally use internal reliability coefficients using the Cronbach's alpha formula (Benek & Akcay, 2019; Korkmaz & Altun, 2014; Sangwan et al., 2021; Warmbrod, 2014). Some researchers found that the highest reliability was on a response scale of 7 to 10 and the lowest reliability was on a response scale of 3 (Taherdoost, 2019). The search results of all reviewed papers show that the use of a Likert scale with 3 response options provides low reliability, which is less than 0.60. The choice category causes some errors in responding to an

attitude phenomenon. In the case of the same study (internal reliability and retest) it became inconsistent. The Likert scale with three response options also results in the loss of a lot of information, especially the intensity and strength of the respondent's opinion (attitude) (Çetin et al., 2020). They also reported that the highest Cronbach alpha coefficient (significant reached 0.96) was achieved on a response scale of 11 and a response scale of 7 with very small differences (James, 2019; Korkut Al Tuna & Arslan, 2016). While the lowest reliability is on the response scale 3. So it can be concluded that the reliability will increase along with the increasing number of response options (from a response scale of 7 to a response scale of 11) even though the reliability is very similar.

Validity

In psychometric tests using several aspects of measurement for validity testing (e.g., Criterion Postdictive Validity, Criterion Predictive Validity, Criterion Concurrent Validity (Benek & Akcay, 2019; Korkmaz & Altun, 2014; Sangwan et al., 2021; Zhang et al., 2021). The validity criteria of the response category show that the response scale 11 is superior to the response scales 3 and 4. According to Preston and Colman, a scale with a response category of 9 has the highest validity (Preston & Colman, 2000). On a scale of 5 to 11 has a very close difference in criterion validity, but increases towards a higher value. Scales with a response category of 6 or more generally have higher convergent validity. Overall, an increase in the number of scale responses will be followed by an increase in validity.

Conclusion

Based on the findings and discussion, the important points that can be conveyed as conclusions are (a) the use of a rating scale with an odd number of answers of more than five points (especially on a seven-point scale) is the most effective in terms of coefficients of reliability and validity, but if the researcher wants to point respondents to one side, then a scale with an even number of responses (six points) may be more suitable; (b) more response choice points (more than five response points) on a Likert scale will increase reliability arguing that the more scores on the scale the higher the level of reliability; (c) overall, an increase in the number of response scales will be followed by an increase in validity; and (e) the presence of response bias and central tendency bias can affect the validity and reliability in the use of the Likert scale instrument.

Recommendations

For future researchers, it is suggested that (a) researchers (reviewers) can establish “quality and relevance” measures to assess the adequacy of each literature review, criteria, and procedures used in collecting data; (b) in the survey, it is better not to use only a closed questionnaire instrument that involves a Likert scale, but an open questionnaire is needed. Open questionnaires can provide the possibility of revealing unexpected things from the respondent's perspective and complement information that is not revealed by closed questionnaires; and (c) We encourage researchers not to worry too much about the number of response choices (Likert scale) in their research. The most important key is to feel comfortable using multipoint items according to the research objectives. However, if you use a questionnaire as a research instrument, we provide some recommendations for compiling and analyzing a survey questionnaire with the aim of minimizing bias and problems that may arise.

1. Think when designing a Likert scale item questionnaire with a balanced key, namely, try to have the number of positive question items equal to negative statement items.
2. Use more question items in the questionnaire and involve a Likert scale with odd response choices (5, 7, 9, or 11).
3. Present the questionnaire (Likert scale) as a bipolar scale and a horizontal presentation.
4. When you are not sure about designing a Likert scale with odd response choices, avoid neutral choices (undecided) and replace them with the words: disagree, have not decided, or other equivalent choices.
5. Various 5-point options were assigned to determine which attitude scale items could be set equivalently as (1) “never”, (2) “rarely”, “(3) “sometimes”, “(4) “often” and (5) "always".
6. The distribution of questionnaires with an online system must include clear instructions for filling out to be more efficient and effective.
7. When filling out the questionnaire through paper and pencil, the respondent must be accompanied if there are questions that the respondent has not understood.
8. The damaged paper and pencil questionnaire results (because there are unanswered question items) are discarded and excluded for analysis.
9. Questionnaires answered by more than 40% of the questions answered "neutral" should be excluded.
10. Questionnaires that are answered in an extreme manner by respondents more than 40% of the question items answered (strongly agree or strongly disagree) should be excluded.

11. Perform data analysis on the results of the questionnaire according to the type of data (ordinal or interval) appropriately and do not analyze the data with an average.

Limitations

This study only reviews literature that has been published in 2012 – 2021. We have not reviewed the latest developments of papers published in early to mid-2022. We hope that researchers who find new developments regarding the use of the Likert scale that have not been discussed in this paper can add new information for future reviews.

Acknowledgements

On this occasion, the researcher would like to thank Sultan Agung Islamic University which has assisted in the financing needed for this research activity.

Authorship Contribution Statement

Kusmaryono: Contributed to the research concept and design, data analysis, interpretation of research results, and final approval. Wijayanti: Data analysis, interpretation, critical revision of manuscripts, manuscript grammar, and supervision. Maharani: Technical support and editing the script.

References

Acosta, S., Garza, T., Hsu, H. Y., & Goodson, P. (2020). Assessing quality in systematic literature reviews: A study of novice rater training. *SAGE Open*, *10*(3), 1–11. https://doi.org/10.1177/2158244020939530

Ahn, E., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean Journal of Anesthesiology*, *71*(2), 103–112. https://doi.org/10.4097/kjae.2018.71.2.103

Aini, Q., Zuliana, S. R., & Santoso, N. P. L. (2018). Management measurement scale as a reference to determine interval in a variable. *Aptisi Transactions on Management*, *2*(1), 45–54. https://doi.org/10.33050/atm.v2i1.775

Alrajeh, T. S., & Shindel, B. W. (2020). Student engagement and math teachers support. *Journal on Mathematics Education*, *11*(2), 167–180. https://doi.org/10.22342/jme.11.2.10282.167-180

Baka, A., Figgou, L., & Triga, V. (2012). "Neither agree, nor disagree": A critical analysis of the middle answer category in Voting Advice Applications. *International Journal of Electronic Governance*, *5*(3–4), 244–263. https://doi.org/10.1504/IJEG.2012.051306

Benek, I., & Akcay, B. (2019). Development of STEM attitude scale for secondary school students: Validity and reliability study. *International Journal of Education in Mathematics, Science and Technology*, *7*(1), 32–52. https://doi.org/10.18404/ijemst.509258

Bidermana, M. D., & Reddockb, C. M. (2012). The relationship of scale reliability and validity to partisipant inconsistency. *Personality and Individual Differences*, *52*(5), 647–651. https://doi.org/10.1016/j.paid.2011.12.012

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International Journal of Exercise Science*, *8*(3), 297–302. https://bit.ly/3ARo13E

Bolarinwa, O. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal*, *22*(4), 195-201. https://doi.org/10.4103/1117-1936.173959

Boone, H. N., & Boone, D. A. (2012). Analyzing Likert data. *Journal of Extension*, *50*(2), Article 2TOT2. https://bit.ly/3RkN2eO

Carey, E., Hill, F., Devine, A., & Szucs, D. (2017). The modified abbreviated math anxiety scale: A valid and reliable instrument for use with children. *Frontiers in Psychology*, *8*(1), 1–13. https://doi.org/10.3389/fpsyg.2017.00011

Çetin, F., Demirkan, Ö., & Çetin, Ş. (2020). A validity and reliability study of the scale for attitude towards classroom as a learning environment. *Educational Policy Analysis and Strategic Research*, *15*(3), 233–248. https://doi.org/10.29329/epasr.2020.270.11

Chen, L.-T., & Liu, L. (2020). Methods to analyze Likert-type data in educational technology research. *Journal of Educational Technology Development and Exchange*, *13*(2), 39–60. https://doi.org/10.18785/jetde.1302.04

Cheng, Y. S. (2012). A measure of second language writing anxiety: Scale development and preliminary validation. *Journal of Second Language Writing*, *13*(4), 313–335. https://doi.org/10.1016/j.jslw.2004.07.001

Çıplak, E., & Çam, S. (2019). The development of the selfie attitude scale: A validity and reliability study. *European*

Journal of Education Studies, *6*(8), 240–254. https://doi.org/10.5281/zenodo.3555247

Dawes, J. (2018). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61–77. https://doi.org/ggktxk

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity*, *52*(4), 1523–1559. https://doi.org/gdqv89

Dilekli, Y., & Tezci, E. (2019). Adaptation of teachers' teaching thinking practices scale into English. *European Journal of Educational Research*, *8*(4), 943–953. https://doi.org/10.12973/eu-jer.8.4.943

Dogan, E. (2018). An application of the partial credit IRT model in identifying benchmarks for polytomous rating scale instruments. *Practical Assessment, Research and Evaluation*, *23,* Article 7. https://doi.org/10.7275/1cf3-aq56

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 364–381. https://doi.org/ckwwnt

Guerra, A. L., Gidel, T., & Vezzetti, E. (2016). Toward a common procedure using Likert and L ikert-type scales in small groups comparative design observations. In M. Dorian, S. Mario, P. Neven, B. Nenad & S. Stanko (Eds.), *Proceedings of the DESIGN 2016 14th International Design Conference* (Vol. 84, pp. 23–32). Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb. https://bit.ly/3Cqvf10

Hartley, J. (2013). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology*, *13*, 83–86. https://doi.org/10.1016/S1697-2600(14)70040-7

James, R. L. (2019). Measuring user experience with 3, 5, 7, or 11 points: Does it matter? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *63*(6), 999–1011. https://doi.org/10.1177/0018720819881312

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, *38*(12), 1217–1218. https://doi.org/b5gxwx

Jeong, H. J., Liao, H. H., Han, S. H., & Lee, W. C. (2020). An application of item response theory to scoring patient safety culture survey data. *International Journal of Environmental Research and Public Health*, *17*(3), 10–14. https://doi.org/10.3390/ijerph17030854

Jeong, J. S., González-gómez, D., & Cañada-cañada, F. (2019). Effects of active learning methodologies on the students' emotions, self-efficacy beliefs and learning outcomes in a science distance learning course. *Journal of Technology and Science Education*, *9*(2), 217–227. https://doi.org/10.3926/jotse.530

Jonnalagadda, S. R., Goyal, P., & Huffman, M. D. (2015). Automating data extraction in systematic reviews: A systematic review. *Systematic Reviews*, *4,* Article 78. https://doi.org/10.1186/s13643-015-0066-7

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975

Józsa, K., & Morgan, G. A. (2017). Reversed items in Likert scales: Filtering out invalid responders. *Journal of Psychological and Educational Research*, *25*(1), 7–25. https://bit.ly/3TLbAze

Khalaf, B. K., & Zin, Z. B. M. (2018). Traditional and inquiry-based learning pedagogy: A systematic critical review. *International Journal of Instruction*, *11*(4), 545–564. https://doi.org/10.12973/iji.2018.11434a

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers and Security*, *64*, 122–134. https://doi.org/10.1016/j.cose.2015.07.002

Korkmaz, O., & Altun, H. (2014). A validity and reliability study of the attitude scale of computer programming learning (ASCOPL). *Mevlana International Journal of Education*, *4*(1), 30–43.

Korkut Al Tuna, O., & Arslan, F. M. (2016). Ölçek madde sayisinin cevaplayicilarin değerlendirmeleri ve veri karakteristiği üzerindeki etkileri: 5'li ve 7 'li likert tipi ölçekler arasindaki farkliliklarin deneysel tasarim kullanarak incelenmesi [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

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items. *Frontiers in Psychology*, *10*, Article 2309. https://doi.org/10.3389/fpsyg.2019.02309

Krosnick, J. A., & Holbrook, A. (2012). The impact of "no opinion" response options on data quality non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, *66*(3), 371–403. https://doi.org/10.1086/341394

Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization

process. *Psychology*, *9*(11), 2531–2560. https://doi.org/10.4236/psych.2018.911145

Lewis, J., & Erdinç, O. (2017). User experience rating scales with 7, 11, or 101 points: Does it matter? *Journal of Usability Studies*, *12*(2), 73–91. https://bit.ly/3bTItIX

Likert, R. (1932). A technique for the measurement of attitudes. In R. S. Woodworth (Ed.), *Archives of Psychology* (Vol. 22, pp. 5–55). SAGE. https://bit.ly/3QngpLX

Lionello, M., Aletta, F., Mitchell, A., & Kang, J. (2021). Introducing a method for intervals correction on multiple Likert scales: A case study on an urban soundscape data collection instrument. *Frontiers in Psychology*, *11*, Article 602831. https://doi.org/10.3389/fpsyg.2020.602831

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

Malone, H., Nicholl, H., & Tracey, C. (2014). Awareness and minimisation of systematic bias in research. *British Journal of Nursing*, *23*(5), 279–282. https://doi.org/10.12968/bjon.2014.23.5.279

Martín, J. C., Román, C., & Gonzaga, C. (2018). How different n-point Likert scales affect the measurement of satisfaction in academic conferences. *International Journal for Quality Research*, *12*(2), 421–440. https://doi.org/10.18421/IJQR12.02-08

Martins, L. E. G., & Gorschek, T. (2016). Requirements engineering for safety-critical systems: A systematic literature review. *Information and Software Technology*, *75*, 71–89. https://doi.org/10.1016/j.infsof.2016.04.002

Mathes, T., Klaßen, P., & Pieper, D. (2017). Frequency of data extraction errors and methods to increase data extraction quality: A methodological review. *BMC Medical Research Methodology*, *17*, Article 152. https://doi.org/10.1186/s12874-017-0431-4

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis* (3rd ed.). SAGE.

Mircioiu, C., & Atkinson, J. (2017). A comparison of parametric and non-parametric Methods applied to a Likert scale. *Pharmacy*, *5*(4), 26–34. https://doi.org/10.3390/pharmacy5020026

Mishra, P., Pandey, C. M., Singh, U., & Gupta, A. (2018). Scales of measurement and presentation of statistical data. *Annals of Cardiac Anaesthesia*, *21*(4), 419–422. https://doi.org/10.4103/aca.ACA_131_18

Mondiana, Y. Q., Pramoedyo, H., & Sumarminingsih, E. (2018). Structural equation modeling on Likert scale data with transformation by successive interval method and with no transformation. *International Journal of Scientific and Research Publications*, *8*(5), 398–405. https://doi.org/10.29322/ijsrp.8.5.2018.p7751

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology*, *44*(1), 369–399. https://doi.org/gg8hfw

Munn, Z., Tufanaru, C., & Aromataris, E. (2014). Data extraction and synthesis. *American Journal of Nursing*, *114*(7), 49–54. https://doi.org/gqbxrm

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *Journal of General Psychology*, *142*(2), 71–89. https://doi.org/gctm2x

Nemoto, T., & Beglar, D. (2014). Developing Likert-scale questionnaires. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings* (pp. 1–8). JALT. https://bit.ly/3AZZqKf

Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. T. (2012). Qualitative analysis techniques for the review of the literature. *Qualitative Report*, *17*(28), 1–28. https://doi.org/gmtqn4

Pedder, H., Sarri, G., Keeney, E., Nunes, V., & Dias, S. (2016). Data extraction for complex meta-analysis (DECiMAL) guide. *Systematic Reviews*, *5*, Article 212. https://doi.org/10.1186/s13643-016-0368-4

Pimentel, J. L. (2019). Some biases in Likert scaling usage and its correction. *International Journal of Sciences: Basic and Applied Research*, *45*(1), 183–191. https://bit.ly/3PwBseJ

Popenoe, R., Langius-Eklöf, A., Stenwall, E., & Jervaeus, A. (2021). A practical guide to data analysis in general literature reviews. *Nordic Journal of Nursing Research*, *41*(4), 175–186. https://doi.org/jbfb

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/dbcr2g

Sangwan, A., Sangwan, A., & Punia, P. (2021). Development and validation of an attitude scale towards online teaching and learning for higher education teachers. *TechTrends*, *65*(2), 187–195. https://doi.org/gjgmqn

Schmidt, L., Olorisade, B. K., McGuinness, L. A., Thomas, J., & Higgins, J. P. T. (2021). Data extraction methods for systematic review (semi) automation: A living systematic review. *F1000 Research*, *10,* Article 401.

https://doi.org/jbfc

Selcuk, A. A. (2019). A guide for systematic reviews: PRISMA. *Turkish Archives of Otorhinolaryngology*, *57*(1), 57–58. https://doi.org/10.5152/tao.2019.4058

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

Sirganci, G., & Uyumaz, G. (2021). Determining the factors affecting the psychological distance between categories in the rating scale. *International Journal of Contemporary Educational Research*, *8*(3), 178–190. https://doi.org/10.33200/ijcer.858599

Solimun, Fernandes, A. A. R., & Arisoesilaningsih, E. (2017). The efficiency of parameter estimation of latent path analysis using summated rating scale (SRS) and method of successive interval (MSI) for transformation of score to scale. *AIP Conference Proceedings*, *1913*, Article 020037. https://doi.org/10.1063/1.5016671

Subedi, B. P. (2016). Using Likert type data in social science research: Confusion, issues and challenges. *International Journal of Contemporary Applied Sciences*, *3*(2), 36–49. https://bit.ly/3q8AVWh

Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/jgme-5-4-18

Taherdoost, H. (2016). Validity and reliability of the research instrument: How to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management*, *5*(3), 28–36. https://doi.org/10.2139/ssrn.3205040

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. *International Journal of Academic Research in Management*, *8*(1), 1–10. https://bit.ly/3Be4KL7

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, Article 45. https://doi.org/10.1186/1471-2288-8-45

Thorpe, G. L., & Favia, A. (2016). Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship*, *20*, 1-33. https://bit.ly/3RcMg39

Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, *50*(6), 2325–2344. https://doi.org/10.3758/s13428-017-0997-0

Ulia, N., & Kusmaryono, I. (2021). Mathematical disposition of students', teachers, and parents in distance learning: A survey. *Premiere Educandum : Jurnal Pendidikan Dasar Dan Pembelajaran*, *11*(1), 147–159. https://doi.org/10.25273/pe.v11i1.8869

Warmbrod, J. R. (2014). Reporting and interpreting scores derived from Likert-type scales. *Journal of Agricultural Education*, *55*(5), 30–47. https://doi.org/10.5032/jae.2014.05030

Xiong, C., Ceja, C. R., Ludwig, C. J. H., & Franconeri, S. (2020). Biased average position estimates in line and bar graphs: Underestimation, overestimation, and perceptual pull. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 301–310. https://doi.org/10.1109/TVCG.2019.2934400

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexao e Critica*, *29*(18), 1–10. https://doi.org/10.1186/s41155-016-0040-x

Zhang, Y., Xu, Q., Lao, J., & Shen, Y. (2021). Reliability and validity of a chinese version of the stem attitude scale for primary and secondary school students. *Sustainability*, *13*(22), Article 12661. https://doi.org/10.3390/su132212661

Zumsteg, J. M., Cooper, J. S., & Noon, M. S. (2012). Systematic review checklist: A standardized technique for assessing and reporting reviews of life cycle assessment data. *Journal of Industrial Ecology*, *16*(1), 12–21. https://doi.org/10.1111/j.1530-9290.2012.00476.x