Available online at www.sciencedirect.com

ScienceDirect

Currents in Pharmacy Teaching and Learning 7 (2015) 836–850

Currents in Pharmacy Teaching & Learning

http://www.pharmacyteaching.com

Review article

How to analyze Likert and other rating scale data

Spencer E. Harpe, PharmD, PhD, MPH*

Midwestern University Chicago College of Pharmacy, Downers Grove, IL

Abstract

Rating scales and rubrics are commonly used measurement tools in educational contexts. Unfortunately, there is a great deal of controversy surrounding how data derived from these tools can and should be analyzed. One issue that is repeatedly raised is whether these data are ordinal or continuous. A related question is whether parametric data analysis techniques are appropriate and/or acceptable for these rating scale data. Some of this controversy may stem from a misunderstanding of fundamental issues related to these particular tools or a poor use of terminology. This article provides a review of basic issues surrounding measurement of various phenomena relevant to educational settings, as well as previous empirical studies examining the effects of using parametric analysis approaches on rating scale data. Based on previous empirical evidence reviewed in this article, parametric analytical approaches are acceptable provided certain criteria are met. Implications for research and teaching are also briefly discussed. After reading this article, the reader should be able to identify the characteristics of a true Likert scale and explain the situations when parametric analytical techniques are potentially appropriate for rating scale data or when nonparametric techniques are preferred.

Keywords: Likert scales; Data analysis; Measurement; Summated scales; Rating scales; Ordinal data

Situation

When I took my first academic position and became involved in educational scholarship, I found myself asked to justify some of my analytical decisions. This was not a bad thing in and of itself since we must be able to justify our choices in any scholarly endeavor. At the time, what struck me as odd were the comments related to how I had analyzed data from rating scales. I remember one passerby at a poster session stopping to look at my poster and saying, “That is interesting, but you know you really should not have used means to report Likert scale results. After all, those are ordinal data.” I was slightly taken aback but thanked the individual for his/her comments and started to think. The statement about Likert scales being ordinal data was never made in my statistics classes in graduate school, so I wrote the comment off as a one-time occurrence. Suffice it to say that was not a one-time occurrence. As a peer reviewer for journals, I sometimes see comments from other reviewers along the same lines as my poster commenter. This issue has also arisen when working with pharmacy practice residents developing their residency projects. After a brief look through the literature, it was clear that the issue of how to analyze data from rating scales and rubrics is not altogether straightforward. The purpose of this article is to provide a review measurement as it relates to the educational context and to provide some evidence-based recommendations for analytical approaches.

Although this article focuses on recommendations for analyzing data from various rating scales for self-reported concepts like self-confidence with a skill learned in class or student perceptions of learning, the discussion herein applies equally to instruments developed to guide external assessment or evaluation (e.g., preceptor evaluation of student performance on an advanced pharmacy practice experience (APPE) or a peer-evaluation of an instructor’s lecture). Nothing in this article is meant to contradict the

* Correspondence to: Spencer E. Harpe, PharmD, PhD, MPH, Chicago College of Pharmacy, Midwestern University, 555 31st Street, Downers Grove, IL 60515.

E-mail: sharpe@midwestern.edu

http://dx.doi.org/10.1016/j.cptl.2015.08.001

recommendations Peeters provides for constructing rubrics to capture rater judgments.[1,2] While that article focuses on the construction of rubrics, the recommendations from this article could be used for the analysis of those ratings.

One word of caution is needed. This article focus on measurement and analysis in the quantitative paradigm that should not be misconstrued as a dismissal of the importance or utility of qualitative research methods. To the contrary, qualitative methods are extremely important and can provide a depth of information that quantitative measurement cannot begin to approach. Information gathered from qualitative methods can inform quantitative methods, and vice versa. Furthermore, both methods can be successfully used together as seen in mixed-methods studies.[3]

Methodological literature review

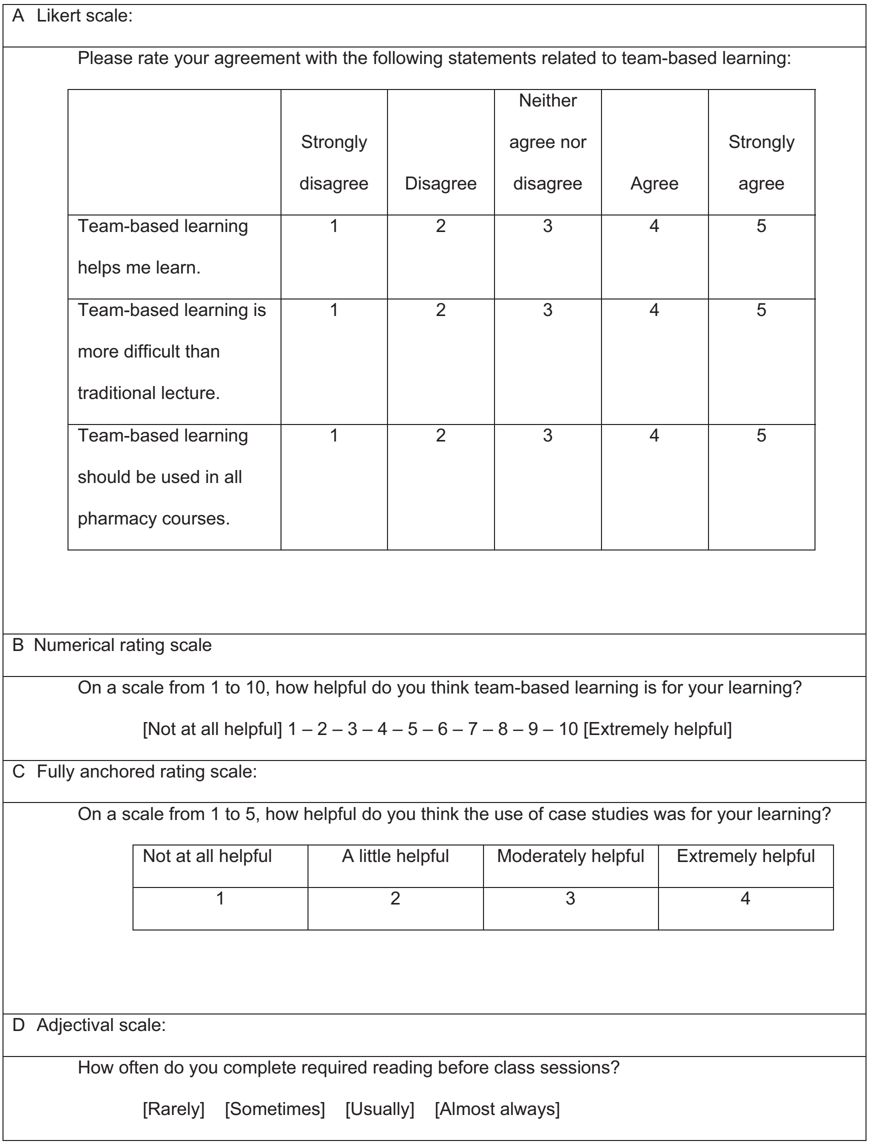
Measurement forms an important part of research and evaluation efforts in the quantitative paradigm insofar as we are concerned with the magnitude, or at least the relative magnitude, of some phenomenon of interest. It is no surprise that the process of measurement is relatively straightforward for physiological quantities (e.g., blood pressure or cholesterol) or for certain clinical outcomes (e.g., hospital length of stay or number of on-time prescription refills). The challenge that we face in education is that many of the phenomena we seek to measure are not physical but are cognitive in nature. The question becomes how can we measure these non-physical phenomena. Some may be directly observable and readily quantifiable, such as the number of times a student commits an error when counseling a patient or the number of seconds an instructor waits for student responses after posing a question during lecture. Other phenomena may be observable yet not directly quantifiable, per se. For example, a preceptor can provide a qualitative assessment of a student's performance of some skill or activity (e.g., "excellent" or "below average"), but a number is not directly observable. Finally, other phenomena, such as self-confidence, are not directly observable at all and must be assessed by self-report. A wide variety of phenomena that are of particular interest in education fall into these latter two categories. Numerous rating scales and rubrics have been developed to allow us to derive quantitative measures of non-physical phenomena by combining a set of items asking an individual to make a series of qualitative assessments. Despite the availability of these instruments, there has been considerable confusion when analyzing data derived from them. The perceived misuse of certain forms of statistical analysis with these instruments has been lamented in the literature,[4,5] yet the recommendations for analysis have stirred considerable controversy.[6–11] The following sections provide a review of the historical development of measurement and associated issues with statistical analysis.

A brief history of measurement

Dating back to the ancient Greeks, measurement is not a recent concept. The modern view of measurement is generally built on Cartesian thinking and was developed during the scientific revolutions of the 15th to 17th centuries. From Descartes' philosophical stance, all *physical* properties could be represented by some quantity. The implication was that quantification was the only option for the scientific method, at least for the natural and physical sciences, thus a quantitative imperative came to be promoted in order for an endeavor to be considered "scientific."[12] This imperative can best be seen when William Thomson (aka, Lord Kelvin) stated that "when you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of an unsatisfactory kind."[13]

The effects of this quantitative imperative on psychological and sociological sciences, and subsequently on educational science, carried into the early 20th century. In the 1930s, the British Association for the Advancement of Science summoned a committee to examine the status of psychophysical measurement methods.[12] In a manner that could be compared to the way a child's performance in school is compared to an older sibling, psychological measurement methods were compared to measurement methods in "older" sciences like physics. The committee generally used Campbell's theory of measurement. This theory described two types of measurements: fundamental measurements and derived measurements.[14] Since derived measurements depended on (i.e., were derived from) fundamental measurements, Campbell viewed fundamental measurements as a necessary step for quantification.[12] Using Campbell's theory, the British committee concluded that psychophysics did not have any "true" measurement methods because there were no fundamental measurement methods. Without a measurement method, psychophysics, and by extension all of psychology, could not be considered a science.[12,15] Simultaneous with the committee's work, and possibly in anticipation of their final report, Stevens set forth a framework for psychological measurement. With this development, psychology could no longer be classified as a non-science since there was now a system of measurement. The implications of Stevens' measurement framework measurement are far-reaching since he proposed the four levels, or scales, of measurement that we continue to use today—nominal, ordinal, interval, and ratio.[16]

In the early-to-mid 20th century, researchers developed various scaling methods to measure the relative strength or degree of psychological phenomena, such as mental abilities and attitudes.[17] As Stevens put rather simply, scaling was a process to assign a number to objects using some rule.[16] Through defining measurement scales, any phenomenon could be "quantified." The development of methods to measure attitudes resulted in several noteworthy scaling



A Likert scale:

Please rate your agreement with the following statements related to team-based learning:

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Team-based learning helps me learn.	1	2	3	4	5
Team-based learning is more difficult than traditional lecture.	1	2	3	4	5
Team-based learning should be used in all pharmacy courses.	1	2	3	4	5

B Numerical rating scale

On a scale from 1 to 10, how helpful do you think team-based learning is for your learning?

[Not at all helpful] 1 – 2 – 3 – 4 – 5 – 6 – 7 – 8 – 9 – 10 [Extremely helpful]

C Fully anchored rating scale:

On a scale from 1 to 5, how helpful do you think the use of case studies was for your learning?

Not at all helpful	A little helpful	Moderately helpful	Extremely helpful
1	2	3	4

D Adjectival scale:

How often do you complete required reading before class sessions?

[Rarely] [Sometimes] [Usually] [Almost always]

Fig. 1. Examples of rating scales and response formats as they may be presented to respondents.

methods including Thurstone scales, Guttman scales, and Likert scales.[18] Likert scales can be included in a larger group of measures that are sometimes referred to as summated (or aggregated) rating scales, since they are based on the idea that some underlying phenomenon can be measured by aggregating an individual's rating of his/her feelings, attitudes, or perceptions related to a series of individual statements or items. Other types of rating scales that are commonly seen include numeric rating scales and adjectival rating scales. Examples of these general types of measurement tools are provided in Figure 1. While there are other measurement options, such as semantic differential scales, visual analog scales, fractionation, and constant sums,[19] these are not frequently seen in educational scholarship and will not be discussed here. It is worthwhile to note the role of Rasch models in measurement. These models provide important alternative to Stevens' approach to scaling (i.e., assigning numbers to objects or categories) and the associated levels of measurement and are being used in educational measurement.[20–22]

Likert[a] scales are a common measurement method in educational contexts.[23] His approach involved the presentation

[a] A discussion of Likert scales would be remiss without addressing another common issue—pronunciation. Many of us probably learned that his name was pronounced as "LIKE-urt." Likert's name is properly pronounced as "LICK-urt"[25 (p45)] so it should more like "sticker" than "hiker."

of a set of selected items that together measured one trait, such as satisfaction with a teaching method. Each item was a declarative statement. For each item, the response set consisted of a set of equally spaced numbers accompanied by approximately equally spaced anchors (Fig. 1A). In the original version, Likert used five options, which included a neutral option.[24] The response format has since been expanded to varying numbers of response options including removing the neutral category.[18,19,25,26] In his original article, Likert proposed that the distances between the numbers in the response set were equal. Similarly, the distances between the anchors (e.g., "Strongly Approve" to "Approve") were equal.[24] From a statistical standpoint, this suggested an interval level of measurement. Rather than analyzing the individual items, Likert combined the individual items via summation or taking the arithmetic mean. By taking the sum or arithmetic mean of the responses to a set of items, Likert scales might arguably be an interval-level measure under Stevens' measurement framework. Likert held that the underlying phenomenon of interest, say satisfaction with teaching, was being measured by an individual's responses to the entire set of items. This is an important point to make. An individual item using a Likert response format (i.e., a Likert item) is *not* a Likert scale.[10,11,24,27,28] Likert scales are a set of items used together. The important difference between the item and the aggregate scale has resulted in a great deal of controversy surrounding "best" or acceptable analytical approaches. Over time, the use of Likert's technique has drifted away from strictly measuring a respondent's level of approval or agreement with statement and may look at frequency (e.g., Never, Sometimes, Often, and Very often) or importance (e.g., Not at all important, Unimportant, Somewhat important, and Very important). Uebersax suggests that these latter examples be referred to as "Likert-type scales" as long as they refer to a set of items that are summed or averaged, are presented horizontally with equally spaced integers, and are presented with labels that are approximately of equal spacing.[28]

Statistical analysis issues

With advances in psychological measurement methods came a great deal of discussion surrounding appropriate methods for analysis. Along with his measurement framework, Stevens also provided a set of "permissible statistics" that could be calculated for data of each measurement level.[16] These are the common descriptive statistics-to-data relationships that many of us were taught in school—modes for nominal data, medians for ordinal data, arithmetic means for interval data, and geometric or harmonic means for ratio data. It is important to note that these permissible statistics are appropriate for any given level of measurement and those levels "above" it in the framework. For example, the arithmetic mean is the recommended statistic for interval data. It is also an acceptable descriptive statistic for data at the ratio level. Similarly, the median is acceptable for ordinal, interval, or ratio data. Stevens also proposed a set of transformations (e.g., multiplication by a constant) and empirical operations (e.g., determination of equality of differences) that could be performed at each level of data. Some took these descriptive statistics as guidelines for the types of inferential statistics that *should* be used. This became known as the Stevens–Siegel–Senders rule given the early textbooks in psychology that promoted this thinking.[29] This rule stated that nonparametric statistical tests were appropriate for ordinal data, and parametric approaches were reserved exclusively for interval or ratio data. Parametric tests were those that assumed the data followed a normal distribution (e.g., *t* test or analysis of variance), while nonparametric approaches were those tests that did not assume a normal distribution (e.g., Mann–Whitney *U* test or Kruskal–Wallis test).[b] This resulted in considerable disagreement within the statistical community. Some argued that the level of measurement was not an assumption of a statistical test.[29] Lord[30] and Wright[31] made similar arguments that the level of measurement is something that researchers assign to the data, rather than the numbers themselves having "knowledge" of their measurement. According to this argument, there was nothing in the data themselves that would prevent us from calculating a mean for data measured at the ordinal, or even nominal, level. From the statistical viewpoint, the distributional behavior of the numbers was what mattered in selecting an appropriate statistical test.[29] On the opposing side of the argument, measurement theorists argued that statisticians were ignoring the importance of measurement theory and that while the numbers may not know how they came to be the researcher did know and must understand this, therefore it was explicitly inappropriate to use parametric statistics for ordinal or nominal data.[32]

For rating scale items, one common criticism is that the distance between two response categories (e.g., the distance from "Strongly disagree" to "Disagree" may not be the same as the distance from "Agree" to "Strongly agree").[18] A similar argument made by Jamieson is that the "average of 'good' and 'fair' is not 'fair-and-a-half.' even when one assigns integers to represent 'fair' and 'good.'"[5] Previous studies looking at the effects of changing measurement levels (e.g., interval to ordinal) or modifying the distance between categories in ordinal measures suggest that these differences are relatively unimportant.[33–37] More recent investigations suggest that when presented with numbers, in either numeric or verbal form or relative magnitude, humans have a mental representation of numbers that seems

[b] An important note related to terminology is warranted here. The data themselves are neither parametric nor nonparametric. Statistical tests can be classified as parametric (e.g., *t* test) or nonparametric (e.g., Mann–Whitney *U* test). Statements like, "The data are parametric," are incorrect; however, one can correctly say, "The data follow a normal distribution."

to resemble a mental number line.[38–41] One proposed phenomenon supporting this mental representation is the spatial–numerical association of response codes, or SNARC, effect whereby under experimental conditions humans responded more quickly to smaller numbers with the left hand and more quickly to larger numbers with the right hand.[39,42,43] This effect appears to be irrespective of "handedness" but is related to the direction of the writing in the subject's native langue (e.g., left-to-right as in English vs. right-to-left as in Arabic).[39] Interestingly, this mental number line has possibly been physiologically mapped onto the brain itself.[44] The mental number line is a particularly important concept for rating scales that provide actual numbers in addition to word anchors. The potential implication is that the intervals between numbers may actually be equal since they appear to be mapped to this internal number line.

Recommendations

A variety of recommendations are currently present in the literature when it comes to analyzing data from rubrics and aggregated rating scales, especially Likert scales. Below are some recommendations about the analysis of data from ratings derived from these summated or aggregated scales, as well as individual rating items. Examples and applications of these recommendations will be provided in the next section. In addition to these recommendations, selected resources for scale development and analysis are provided in Figure 2.

Before making recommendations, some terminology issues deserve a brief discussion. Unfortunately, rating scale terminology is used inconsistently, especially the term "scale."[10] This is problematic as it results in a lack of clarity surrounding the intended meaning. Research methods texts frequently define a scale as a group of items.[17,25] At times a scale may refer to individual item, such as a numerical rating scale.[45] The term can even be used to refer to the type of data, as in Stevens' scales of measurement.[16] Unfortunately, it is difficult to make blanket statements about what the term "scale" means when seen in a published study since the use of the term may be context dependent. For the purposes of this article, the distinction between an aggregated scale composed of multiple items and an individual item will be made as clear as possible.

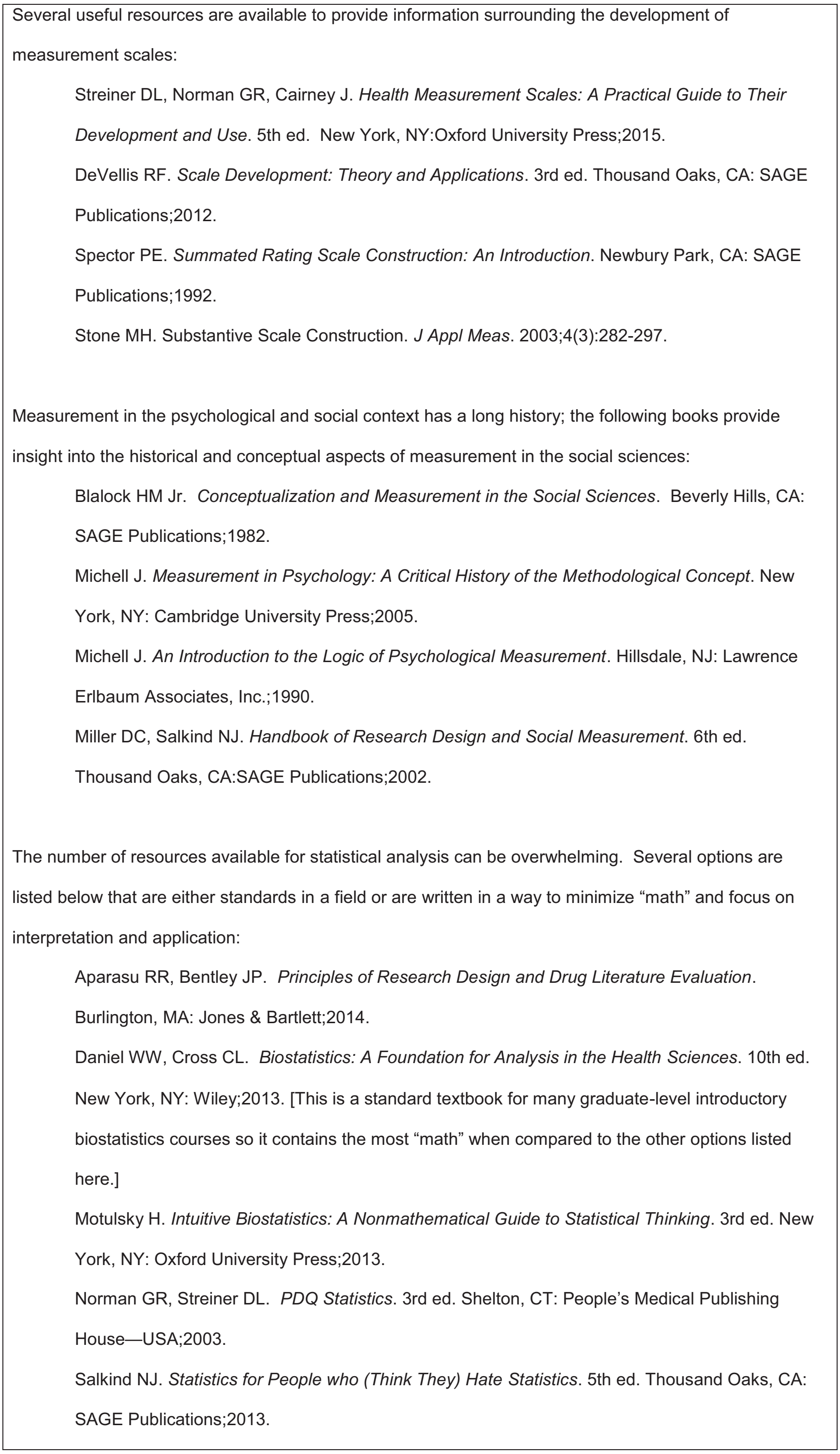
Recommendation 1: Scales that have been developed to be used as a group must be analyzed as a group, and only as a group

This first recommendation is of utmost importance and addresses some of the confusion that appears to have been promoted by those espousing "ordinalist" views on aggregated scales.[4,5] When aggregated rating scales like Likert scales are developed, the initial psychometric evaluation examines the performance of those items *as a group*. When these aggregated scales are used in a study, they *must* be analyzed as a group. If the scale contains validated subscales, then these may be examined separately since the development of the scale supported the presence of those subscales.[10,46,47] The individual item-by-item analytical approach, however, is inappropriate for methodological and statistical reasons as discussed in the following paragraphs.

From a measurement standpoint, the individual item is not a measure of the overall phenomenon of interest. Drawing on Likert's original thinking, the phenomenon of interest is measured by the aggregate group of items in the scale, not simply by any one item on its own. Separating the items conceptually "breaks" the theoretical measurement properties of the aggregated scale as it was originally developed. This might be likened to the ecologic fallacy in epidemiologic research where the conclusions at the group level do not necessarily apply to the individual level.[48] One way to avoid this is to make the unit of analysis reflect the unit of inference. For example, when analyzing hospital-level antibiotic use and resistance data, the inferences and conclusions must also be at the hospital level, not the individual patient level unless some sort of multi-level analysis was performed. In a similar fashion, it would not be prudent to analyze class- or program-level data and then make inferences about expectations for an individual student's outcomes. Putting this back into a measurement context, the measurement approach and the analytical approach must be aligned. If the phenomenon is measured by an aggregated score on a set of items, then that aggregated score *must* be used for analysis.

Analyzing the items one-by-one also causes statistical concerns through an inflation of the family-wise error rate. This is similar in concept to the use of multiple comparisons tests after an analysis of variance (ANOVA). There are various methods used to ensure that the overall Type I (α) error rate is maintained at the stated level (e.g., 5%) including the common Bonferroni procedure or the more powerful Holm procedure.[49] Whether this inflated Type I error rate is of practical concern is the source of debate. Some argue that adjustment for multiple comparisons to avoid inflated Type I error rates is not necessary, especially in exploratory contexts.[50,51] Perneger[52] argues that the statistical significance of one finding does not conceptually (or philosophically) depend on whether another finding is determined to be statistically significant. In cases where many individual items are being analyzed (e.g., comparing two groups of students on 20 separate items, *not* some aggregate score of the 20 items), it may be worthwhile to consult with a statistician colleague to determine whether any adjustment for multiple comparisons is needed and how to make the adjustments.

There are some situations where item-by-item analyses may be acceptable; however, these are very few. Items may be examined individually during instrument development, especially when a new instrument is being compared to an



Several useful resources are available to provide information surrounding the development of measurement scales:

Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 5th ed. New York, NY:Oxford University Press;2015.

DeVellis RF. *Scale Development: Theory and Applications*. 3rd ed. Thousand Oaks, CA: SAGE Publications;2012.

Spector PE. *Summated Rating Scale Construction: An Introduction*. Newbury Park, CA: SAGE Publications;1992.

Stone MH. Substantive Scale Construction. *J Appl Meas*. 2003;4(3):282-297.

Measurement in the psychological and social context has a long history; the following books provide insight into the historical and conceptual aspects of measurement in the social sciences:

Blalock HM Jr. *Conceptualization and Measurement in the Social Sciences*. Beverly Hills, CA: SAGE Publications;1982.

Michell J. *Measurement in Psychology: A Critical History of the Methodological Concept*. New York, NY: Cambridge University Press;2005.

Michell J. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.;1990.

Miller DC, Salkind NJ. *Handbook of Research Design and Social Measurement*. 6th ed. Thousand Oaks, CA:SAGE Publications;2002.

The number of resources available for statistical analysis can be overwhelming. Several options are listed below that are either standards in a field or are written in a way to minimize “math” and focus on interpretation and application:

Aparasu RR, Bentley JP. *Principles of Research Design and Drug Literature Evaluation*. Burlington, MA: Jones & Bartlett;2014.

Daniel WW, Cross CL. *Biostatistics: A Foundation for Analysis in the Health Sciences*. 10th ed. New York, NY: Wiley;2013. [This is a standard textbook for many graduate-level introductory biostatistics courses so it contains the most “math” when compared to the other options listed here.]

Motulsky H. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. 3rd ed. New York, NY: Oxford University Press;2013.

Norman GR, Streiner DL. *PDQ Statistics*. 3rd ed. Shelton, CT: People’s Medical Publishing House—USA;2003.

Salkind NJ. *Statistics for People who (Think They) Hate Statistics*. 5th ed. Thousand Oaks, CA: SAGE Publications;2013.

Fig. 2. Recommended resources for scale development and statistical analysis.

existing instrument, or when an item-by-item analysis is the best that can be done. This may happen when a new set of items is constructed in some meaningful fashion but no psychometric evaluation has been conducted. In this situation, the indefinite use of the set of items without psychometric evaluation is problematic and should be avoided if at all possible. One could also make an argument for the presentation of the responses for each item, such as with diverging stacked bar charts as proposed by Robbins and Heiberger,[53] as long as formal inference at the item level was not performed. Generally speaking, however, the item-by-item analysis of aggregated scales should be avoided.[10]

Recommendation 2: Aggregated rating scales can be treated as continuous data

Some view Likert scales as being strictly ordinal in nature, thus parametric analysis approaches assuming quantitative, or at least interval level, measurements are not appropriate.[4,5,9] Carifio and Perla[11] suggest that this "ordinalist" view stems from Stevens' argument that phenomena which are ordinal at the individual level (e.g., a Likert item) cannot be interval (or ratio) at the aggregate level (e.g., a Likert scale). Such ordinalist tendencies can be found in some well-respected textbooks and reference resources. For example, Miller and Salkind state that "the Likert technique produces an ordinal scale that generally requires nonparametric statistics."[45 (p330)] Cohen, Manion, and Morrison are more forthright in their position: "Rating scale questions are treated as ordinal data, though one can find very many examples where this rule has been violated, and nonparametric data have been treated as parametric data. This is unacceptable."[18 (p390)] Many of these "ordinalist" views seem to stem from overlooking the difference between individual items and overall Likert scales as pointed out by those holding the contrasting "intervalist" viewpoint.[8,10,11,23] This intervalist vs. ordinalist argument can be extended to other aggregated rating scales.

The idea of banning the use of parametric analyses for aggregated scale data may be an overly restrictive approach in practice since numerous studies have suggested that parametric approaches are acceptable when the data are not strictly on the interval level of measurement.[33–37] Many of the methods used to develop and validate these scales (e.g., factor analysis and item response theory) are based on parametric statistics and use categorical variables to create a measure of an underlying latent variable, which are considered continuous variables.[46,54] With this in mind, the generalist calls to use only nonparametric approaches for aggregated rating scales is overly restrictive and not reflective of the manner in which these scales were developed.

This recommendation does come with some caveats. While parametric analysis approaches *may* be appropriate, verifying the assumptions of a particular statistical test is still important since violated assumptions, especially deviations from normality, can result in a loss of statistical power (i.e., the probability of identifying a difference or effect when one actually exists).[55] Common assumptions include independence of observations, homogeneity of variance, and a normal distribution. The independence assumption must be verified conceptually, and alternate tests should be used when this is violated (e.g., the paired *t* test). Few would disagree that perfect normal distributions are rarely found in practice. Similarly, perfectly equal variances between groups (i.e., homogeneity of variance) are exceedingly rare. In practice, we are faced with determining whether a distribution is "normal enough" or the variances are "equal enough." Normality can be readily assessed through graphical means, such as histograms, density estimates, and normal Q–Q plots. Equality of variance can be crudely assessed by comparing the group standard deviations. There are also formal statistical tests to assess normality (e.g., Kolmogorov–Smirnov or Shapiro–Wilk tests) and equality of variance (e.g., Levene's test, Bartlett's test, or the Brown–Forsythe test), which are implemented in most common statistical packages. Thankfully, empirical studies have shown that some common parametric methods, especially the *t* test and *F* test, are relatively robust to violations of normal distribution and equality of variance provided that two-tailed hypothesis tests are used and the sample sizes within groups are reasonably similar.[33,36] We can also draw comfort from the central limit theorem since the distribution of the sample mean or differences in the means will be normally distributed with reasonably large sample sizes.[23] Still, it is important to keep in mind situations where a nonparametric analysis approach may be preferred: a one-tailed test is genuinely needed, the group sizes are substantially different, sample sizes are moderate, or there is a severe departure from normality.

Recommendation 3: Individual rating items with numerical response formats at least five categories in length may generally be treated as continuous data

Aside from aggregated rating scales, it is not uncommon to see individual rating items used. These may appear as individual Likert items outside of a composite scale or a holistic rubric score made by external raters (e.g., a faculty member judging a student's performance in a simulation; Peeters discusses important points to consider when developing these rubrics[1,2]. A common example is when a student is asked to rate their satisfaction with a teaching method when provided with a numerical scale (e.g., 1 = "Not at all satisfied" and 10 = "Completely satisfied"). Johnson and Christensen[54] provide a typology for these individual items. When anchors or labels are only provided at the extremes, these can be called numerical rating scales (Fig. 1B). Fully anchored rating scales represent the situation when an anchor is provided for each numerical option (Fig. 1C). An individual Likert item could be viewed as a specific instance of a fully anchored rating scale where the anchors represent levels of agreement. Single-item measures may be appealing from the perspective of brevity and may even be necessary because of the lack of availability of more formally developed aggregated scales or the secondary use of existing data. Aside from carefully constructed holistic rubrics as described by Peeters[1,2], single-item measures should be used with caution since their use is problematic given the fact that psychometric evaluation is difficult and sometimes impossible.[47]

The response format for these individual rating items is often presented as some sort of number line. Previous research examining numerical processing supported the use of a mental number line when numerals were presented, but similar effects were not seen when verbal numbers or words were used in the experiments.[38,39] This suggests that the use of actual numbers in the presentation of the rating scale may be useful by potentially allowing us to treat responses as interval-level measures rather than simply as ordinal data. This is why non-numerical response formats, such as the adjectival rating scale described by Streiner, Norman, and Cairney[26] are included below in Recommendation 4.

Research has shown that the number of categories in a response format, the word anchors used for the response categories, and the numerical representation (e.g., 0 to 10 vs. −5 to 0 to +5) are all important factors affecting how an individual responds to an item.[56–58] While these may affect the shape of the resulting distribution or engender certain biases within respondents, the responses still appear to be somewhat interval in nature given the use of the numeric presentation and consideration can be given to parametric analytical methods. Hsu and Feldt[59] found that with at least five categories, the use of ANOVA was appropriate even for considerable differences in variance across groups. The caveats about normality and equality of variance provided in Recommendation 2 should be also considered here.

One extremely important point related to data distribution bears a brief discussion. Although an item may have been developed with the intention that respondents might use all possible categories (e.g., from 1 = "Strongly Agree" to 5 = "Strongly Disagree"), respondents may only use two or three of the available response categories because of the wording of the item or the nature of group being studied. This would result in a "shorter" item that would fall under Recommendation 4. The determination between whether to follow Recommendation 3 or Recommendation 4 should be made after looking at the actual distribution of the responses (i.e., whether the full range of response options used or only a few).

Recommendation 4: Consider nonparametric or categorical data analysis approaches for individual rating items with numerical response formats containing four or fewer categories or for adjectival scales

Special consideration is needed when the number of response categories become relatively small or when actual numbers are not provided for responses and only a set of ordered category descriptions are provided (referred to by Streiner, Norman, and Cairney as adjectival scales[26]; Fig. 1D). In these situations, the distribution of the responses may be difficult to justify as being at least interval-level measurements. Also, previous statistical investigations may have suggested a substantial concern for analytical problems associated with unequal variance or large levels of skewness.[59,60] In these cases nonparametric approaches would be appropriate. In cases where meaningful collapsing can occur (e.g., collapsing to agree vs. disagree or pass vs. fail), categorical approaches may provide a better alternative from an interpretation standpoint. Nonparametric approaches like the Mann–Whitney *U* test or the Kruskal–Wallis test look for differences in the distributions of data between groups. For example, a Mann–Whitney *U* test could be used to examine whether the distribution of students' ratings of how frequently they review their notes after class differs between two groups of students (rated as 1 = Never, 2 = Rarely, 3 = Sometimes, and 4 = Usually). The test does *not* answer whether the mean or median rating is different between the groups of students. Technically, categorical tests, like the Chi Square test, also examine differences in distribution of responses; however, categorical tests are often framed as looking at differences in percentages between groups. Concluding that the percentage of students who agree with a given statement is higher in one group than another might be more meaningful than a more general, but no less correct, conclusion that the distribution of the responses is different between the groups.

Recommendation 5: Remember the benefits of statistical models

Simple group comparison via *t* tests or ANOVA is useful, but it is important to mention that more advanced modeling strategies such as linear regression or analysis of covariance (ANCOVA) may also appropriate be for use with continuous data since they allow the consideration of other covariates and confounders in the analysis. These methods come with their own assumptions that must be verified in order for the results to be valid statistically. Thankfully there are alternatives when assumptions are violated. For example, linear mixed models can be used when the data are approximately normally distributed but the observations are not independent (e.g., due to matching of students or nesting arising from groups of students being in the same group, class, or institution). When distributional assumptions are not met, the modeling approaches can get more complicated. If the data can be meaningfully categorized, then categorical models such as logistic regression can be used. When categorization is not ideal, then a generalized linear model can be used where the approximate data distribution is specified in the model. It may be beneficial to consult colleagues knowledgeable about these

Table
Outcome measures for Harpe et al.[61]

Questionnaire	Description	Measurement and scoring	Example items from the questionnaire
Self-efficacy to learn statistics (SELS)	Student's self-reported self-efficacy to learn statistical concepts	14 Items rated from 1 (no confidence at all) to 6 (complete confidence) Scoring: Sum of the items (minimum: 14, maximum: 84) Higher scores indicate higher self-efficacy	Identify scale of measurement for a variable Identify the factors that influence power Identify when the mean, median, and mode should be used as a measure of central tendency
Current statistics self-Efficacy (CSSE)	Student's self-reported current statistical self-efficacy (similar to SELS but focuses on performance rather than learning)	14 Items rated from 1 (no confidence at all) to 6 (complete confidence) Scoring: Sum of the items (minimum: 14, maximum: 84) Higher scores indicate higher self-efficacy	Same items as SELS but instructions prompt respondent to focus on self-confidence in performance rather than learning
Survey of attitudes towards statistics (SATS-36)	Self-reported attitudes towards statistics in general; six domains: affect (feelings towards statistics), cognitive competence (attitudes about their intellectual knowledge), value of statistics, difficulty, interest in statistics, effort put forth in the statistics course; one version for pre-course assessment and different version for post-course assessment (same core items but worded to be consistent for use after completion of the course)	36 Items rated from 1 (Strongly disagree) to 7 (Strongly agree) Scoring: Mean score for each domain (minimum: 1; maximum: 7) Higher scores indicate more positive attitudes (some items are reverse coded in calculating the mean domain score)	I will like statistics I will find it difficult to understand statistical concepts Statistics should be a required part of my professional training Learning statistics requires a great deal of discipline I am interested in understanding statistical information I plan to attend every statistics class session
Student perceptions of the course	Students' perceptions of various course attributes that are modified in learning-centered courses; three areas: course structure and activities, course policies, preference for the	13 Items rated from 1 (Strongly disagree) to 5 (Strongly agree) Scoring: mean score for each item; also percentage in agreement with each item calculated using those responding "Agree" or "Strongly agree"	I was provided with multiple opportunities to demonstrate that I had learned the material Course policies supported my learning in the course If given the option, I would rather take a drug literature evaluation course using a learner-centered approach than a drug literature evaluation course with a more "traditional" approach

Modified and reprinted with permission from Harpe et al.[61]

more advanced approaches when statistical modeling procedures are being used.

Applications

Two articles from pharmacy education have been chosen to illustrate the previous recommendations in action. The first is a study of the effect of learning-centered assessment on statistical knowledge and attitudes.[61] For this study, the authors used several measures for the study outcomes (Table). Two measures were related to statistical self-efficacy: self-efficacy to learn statistics (SELS) and current statistics self-efficacy (CSSE).[62] Both of these instruments consisted of only one domain. Attitudes towards statistics were measured by the 36-item Survey of Attitudes Towards Statistics (SATS-36), which consisted of four validated domains: affect, cognitive competence, value, and difficulty.[63] The SELS, CSSE, and SATS-36 are all Likert scales. Previously developed instruments were used to measure statistical knowledge[64] and student course perceptions.[65]

For the measures of statistical attitudes, self-efficacy, and knowledge, the instruments were scored as described by the instrument developers and analyzed only as aggregate scores as mentioned in Recommendation 1. For the SATS-36, this involved analyzing the data as the four distinct subscales rather than an overall score, as directed by

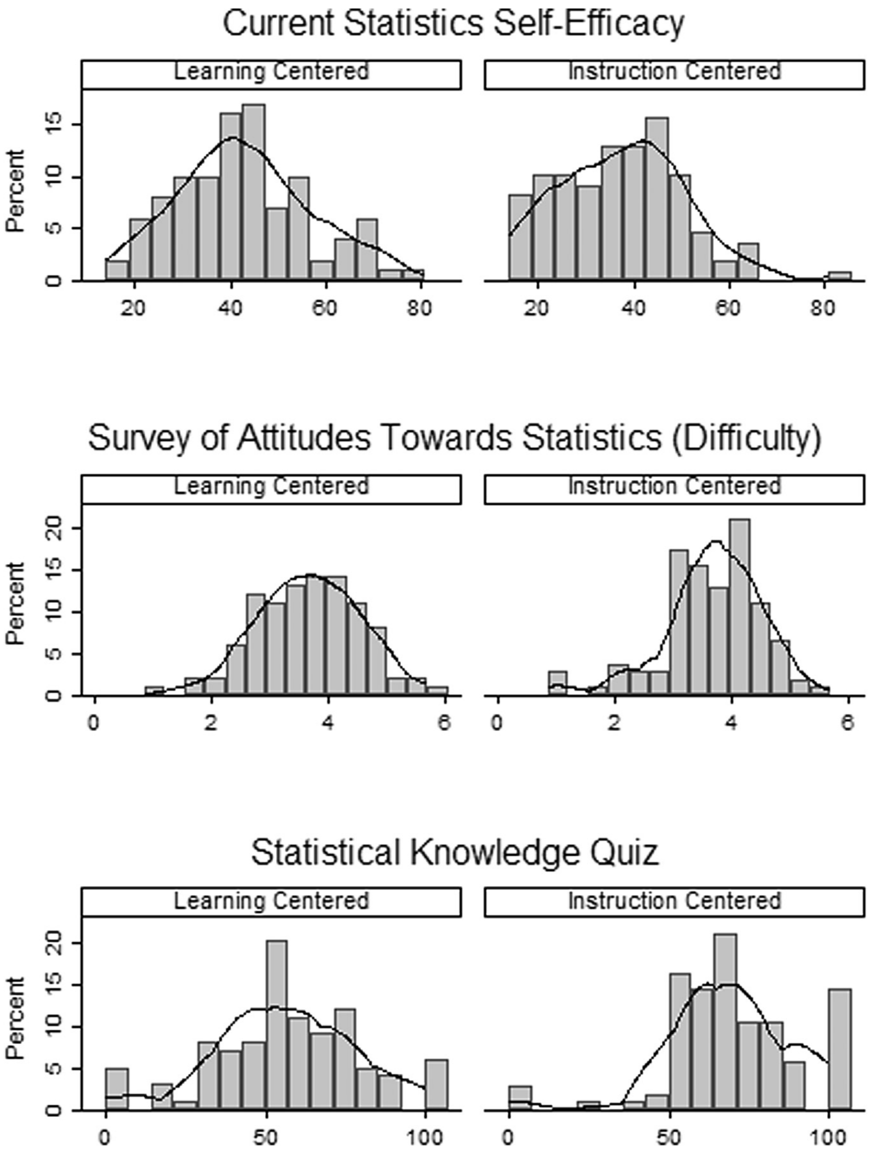


Fig. 3. Distributions of selected pre-test measures for Harpe et al.[61] The bars represent a histogram and the solid line is an overlay of a density estimate for each measure to provide an alternate view of normality. CSSE = Current Statistics Self-Efficacy scale; SATS-36 = Survey of Attitudes Towards Statistics, 36-item scale.

the instrument developer.[63] It is important to remember that scores from a subset of items cannot be created unless the reliability of that subset has been demonstrated. The individual items for the course perceptions were analyzed individually since the psychometric properties of the instrument were unknown. The three measures for attitudes and self-efficacy were treated as continuous variables as noted in Recommendation 2. Figure 3 shows a histogram with a density estimate overlay for the pre-test of three of the measures used in the study to demonstrate the graphical assessment of normality. While none of the distributions are perfectly normal, the graphs do show that a certain level of normality can been assumed without providing substantial reason to worry about violating assumptions required for parametric approaches given the robustness of these procedures discussed earlier. It is important to note that these graphs need not be included in a published article. What is important is that the distributional assumptions are examined and considered when selecting a data analysis technique.

The analysis of student course perceptions illustrates issues associated with Recommendations 3. Each course perception item was measured using a five-point Likert response format (1 = Strongly Disagree to 5 = Strongly Agree). Since the response format contained at least five categories, the authors treated these as continuous variables and reported means and standard deviations for each item. Similarly, each item was compared between the learning-centered assessment group and the traditional assessment group using a parametric approach (unpaired *t* test). For added support of the findings, the authors also analyzed these items using a nonparametric approach (Mann–Whitney *U* test) and collapsed the data into two groups (Strongly Agree/Agree vs. Neutral/Disagree/Strongly Disagree) to use a categorical approach (Fisher's exact test). If the student perceptions items had only used four response categories (e.g., Strongly Agree, Agree, Disagree, and Strongly Disagree), then the collapsing could have been easier (as mentioned in Recommendation 4). The analysis of the course perception items does, however, highlight a less than desirable situation. Although the course perception items were used previously, there was no psychometric evaluation. It is possible that an aggregated scale could have been created from these items, but this is currently unknown without more formal scale development techniques being applied to the data. Furthermore, there were 13 comparisons being made, which may have resulted in an effective Type I error rate greater than the nominal 5%. While the item-by-item analysis for course perceptions may have been acceptable in the short term since the items are relatively new, future use may not be recommended without further evaluation of those items.

As a brief second example, consider the study by Huynh et al.[66] examining readiness for self-directed learning before and after an advanced pharmacy practice experience (APPE). The main measure was the Self-Directed Learning Readiness Scale (SDLRS) developed by Fisher et al.[67] The authors analyzed the SDLRS results as a whole (Recommendation 1) by following the scoring set forth by Fisher et al. including the use of the three SDLRS subscales. For the sake of completeness, the SDLRS scores were described with the mean and the median. In comparing the pre- and post-APPE SDLRS scores, the authors used a paired *t* test approach (Recommendation 2). Similar to the Harpe et al. study, the authors supplemented the parametric analysis with a categorical approach using McNemar's test to examine whether students scored over 150 on the SDLRS before and after APPEs. Similar conclusions were reached with both analytical approaches. The only area of concern was the description of the SDLRS as using a five-point "Likert scale."[66] From a clarity standpoint, it is important to avoid confusing the Likert response format with the overall Likert scale. The way the authors worded the statement it suggests that the fact the SDLRS uses five response categories to measure agreement makes this a Likert scale. This is not technically correct. The fact that the SDLRS uses a *set* of (psychometrically validated) items where respondents rate their level of agreement with each item and it is presented in a particular manner is what makes this a Likert scale. If the SDLRS consisted of only one item, it would not be a Likert scale even if the respondent reported her level of agreement. A better description could have been that the SDLRS was a 40-statement Likert scale. This clarifies that the Likert scale comes from the aggregated items rather than the five-point agreement response scale.

Implications

In light of the recommendations made in this article, there are some important implications for our individual practices. These will be loosely separated by research design and implementation, research reporting, and teaching. Obviously, these are not completely independent areas, so there may be some conceptual overlap. One overall implication is the imperative for us to be clear and consistent in what we do with respect to the use of rating scales.

Research design and implementation

When designing a study, the most important step is defining the research question or the research objective since all other steps flow from that starting point. In developing the measurement methods for a study, one common recommendation is to use previously developed scales or modify these scales whenever possible.[46,68] Using pre-developed instruments is often a good start. We must remember to provide important psychometric assessments in our local usage, such as the reliability of the instrument in our particular sample, as recommended by Hoover et al.[69] Unfortunately, the identification of a previously developed instrument with unknown psychometric properties can and does happen. The lack of psychometric evaluation can raise the questions of reliability and validity of the instruments. In an effort to follow the developers' methods, researchers sometimes engage in less than optimal strategies, such as an item-by-item analysis. Performing such analysis simply because the developer did so only perpetuates less than optimal analytic strategies and may lead to incorrect conclusions. At some point, these existing but previously non-validated instruments must undergo psychometric evaluation or a new, validated instrument should be developed. This also applies to instruments that we must develop ourselves because no appropriate ones exist. A related situation is the modification of previously developed validated instruments. Modifications should be avoided if at all possible since these may alter the established psychometric properties of the instrument. When modifications are absolutely necessary, re-evaluation of the psychometric properties is needed.

The choice of a statistical test is ultimately context specific and must be justifiable. The choice should be guided by the research question or objective in relation to the measurement methods. In some situations, a nonparametric test may be required or preferred. For example, assume that we are comparing two groups of students with respect to their satisfaction of a certain teaching strategy. Satisfaction is measured with an item where students provide a rating from 1 (Not at all Satisfied) to 5 (Completely Satisfied). We established earlier that parametric alternatives could be acceptable to use for this analysis. Let us now assume that the students in the two groups have the same mean satisfaction; however, in one group, students appear to be polarized such that they move toward the two extremes of the response options and the distribution of their satisfaction ratings is bimodal. Despite having the same means, there are two distinctly different distributions on the satisfaction rating. A categorical test (e.g., Chi Square test) or nonparametric tests that compare distributions (e.g., Mann–Whitney *U* test) would detect this difference in distributions but a comparison of means (e.g., *t* test or ANOVA) might not.[59] There are also situations where parametric approaches may be preferred over nonparametric options, such as when interactions are needed or when examining more than two variables simultaneously (e.g., regression models). In these situations, there are alternatives to the traditional least squares regression model that can take into account the distribution of the data rather than assuming a normal distribution. Making overly prescriptive commands about which statistical tests are "right" or "wrong" or that certain tests are "not permissible" is not helpful since these can ignore the nuances of a particular data set and the process of data analysis. This is why the previous recommendations are just that—recommendations. Ultimately, selecting the "best" statistical test should be related to knowledge of our actual data and the overall research question.

Research reporting

When reporting our results, it is important to use language that clarifies the difference between a summated or aggregate scale, an individual rating item, and a response format. Some have argued that the lack of clarity in reporting has led to the current state of confusion surrounding appropriate analytical strategies.[10,28] It is imperative that appropriate terminology is used in reporting. In particular, a Likert scale (or a Likert-type scale) should only refer to a set of items.[8,10,24,28] If being used appropriately (i.e., not outside of the aggregated Likert scale), an individual item with a Likert response format might appropriately be termed a "Likert item" or a "Likert-type item" if it is being used individually. There are certain situations where the term "rating" is not even appropriate. Consider an item asking students how many hours per week they study with the following response set: 0 = Less than 2 hours; 1 = 2 to 4 hours; 3 = 5 to 7 hours; 4 = 8 to 10 hours; 5 = 11 or more hours. This is not a scale nor does it involve "rating." It is simply an ordered response category.

Another implication for reporting is the issue of item-by-item analyses. With aggregated scales there may be a desire to see if there are any differences in the component items of the scale. Drilling down to the individual item level violates the central measurement concept of the instrument since it was developed and validated as a whole. If the instrument has component subscales, then a *post hoc* comparison of those subscale scores could be warranted depending on the guidelines provided by the instrument developers. An example without an aggregated scale might provide another point of view. Consider an instructor interested in evaluating whether changing a teaching strategy affects knowledge as measured by the overall course grade for students. The method of teaching is changed for students in the current academic year with students in the previous year serving as a comparison group. For this study, the major assessment methods in the course remain the same and the exact same examinations are used. If the actual evaluation question were about the differences in overall knowledge defined as final course grade, then looking at the individual test grades might not be meaningful. If certain groups of topics could be mapped to certain exams, then looking at the tests individually would allow her to examine whether the teaching method was more useful for certain topics than others. Comparing performance between the two classes (current year vs. previous year) on each test question would not be appropriate since the course grade as a whole was taken to measure knowledge. This question-by-question analysis is analogous to dismantling a validated aggregated scale for item-by-item analysis.

In situations where a new set of items has been developed but not evaluated psychometrically, there may be some room for item-by-item analysis. The items should be examined for their psychometric properties if they will be used in later studies, though evidence of reliability from your sample and content validity for your instrument should suffice until a more thorough analysis is done.[70] At a minimum, the limitations associated with not knowing those psychometric properties should be clearly stated when reporting the results.

Teaching

From a teaching standpoint, we need to do a better job of helping students understand some of the nuances associated with measurement, especially related to rating scales. Statements like, "Likert scales are ordinal data, so parametric statistics are not appropriate," may be used in order to simplify statistics for students. In reality, however, overgeneralizations like this only contribute to confusion as they do not always make the distinction between individual items and summated or aggregate scales. Of somewhat greater concern, these overgeneralizations ignore past research that speaks directly to the analysis of these types of data. As someone who has taught research methods and statistics to pharmacy students, I can fully understand the difficulty in helping students learn material that is important to their ability to be evidence-based practitioners while keeping things as simple as possible so that they do not get lost in some of the more esoteric details. Simplification may be useful to help students understand certain basic principles; however, we must ensure that we introduce students to greater levels of complexity in order to avoid doing them a disservice. For example, we would never make a blanket statement such as, "All patients with diabetes are treated with insulin." It is true that *some* individuals with diabetes will require insulin therapy, but an oversimplification like the previous statement would be incorrect and could potentially cause harm.

Another implication for teaching that relates to the idea of consistency is the discussion of various other aggregated scales that students may be introduced to throughout their training. These may include the Center for Epidemiologic Studies Depression Scale (CES-D)[71] or the Medical Outcomes Study short-form measures of health like the SF-36.[72] Rarely, if ever, do people balk at the use of parametric statistical approaches for these measures, both of which represent aggregated scales where individual items might technically be considered ordinal levels of measurement since they are Likert-type items. It seems that the presence of the term "Likert" has become a heuristic that is not necessarily correct. Perhaps one final implication is that we take time to educate ourselves on the topic of the differences between items and true scales.

Conclusions

The controversy surrounding the appropriate analysis of various types of rating scales has existed for over 65 years dating back to the time when the original framework for levels of measurement was proposed by Stevens. Over time a wide variety of studies have been conducted examining the statistical properties of data from these summated or aggregate scales and from individual items with various forms of rating. The general consensus is that the use of parametric statistical tests to analyze these data is potentially appropriate taking into consideration the recommendations provided here. While this one article is not expected to end the controversy, it may serve to provide support and offer some clarity to those who are using these common measurement tools.

Conflicts of interest

The author has no conflicts of interest to disclose with respect to the authorship and/or publication of this article.

References

1. Peeters MJ. Measuring rater judgments within learning assessments, Part 1: why the number of categories matter in a rating scale. *Curr Pharm Teach Learn*. 2015;7(5):656–661.
2. Peeters MJ. Measuring rater judgments within learning assessments, Part 2: a mixed approach to creating rubrics. *Curr Pharm Teach Learn*. 2015;7(5):662–668.
3. Creswell JW. *Research Design: Qualitative. Quantitative, and Mixed Methods Approaches*. 4th ed., Thousand Oaks, CA: SAGE Publications; 2013.
4. Kuzon WM. Jr, Urbanchek MG, McCabe S. The seven deadly sins of statistical analysis. *Ann Plast Surg*. 1996;37(3):265–272.
5. Jamieson S. Likert scales: how to (ab)use them. *Med Educ*. 2004;38(12):1217–1218.
6. Armstrong GD. Parametric statistics and ordinal data: a pervasive misconception. *Nurs Res*. 1981;30(1):60–62.
7. Knapp TR. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nurs Res*. 1990;39(2):121–123.
8. Pell G. Use and misuse of Likert scales. *Med Educ*. 2005;39(9):970.
9. Jamieson S. Use and misuse of Likert scales—author's reply. *Med Educ*. 2005;39(9):971–972.
10. Carifio J, Perla R. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *J Soc Sci*. 2007;3(3):106–116.
11. Carifio J, Perla R. Resolving the 50-year debate around using and misusing Likert scales. *Med Educ*. 2008;42(12):1150–1152.
12. Michell J. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1990.
13. Thomson W. *Popular Lectures and Addresses*, vol. 1. London, UK: Macmillan Publishers; 1891, p. 80–81.
14. Campbell NR. *Physics. The Elements*. Cambridge, UK: Cambridge University Press; 1920.
15. Bartlett RJ. Measurement in psychology. The Advancement of Science: the Report of the British Association for the Advancement of Science. 1:422–441; 1939-1940.
16. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103(2684):677–680.
17. Singleton RA. Jr, Straits BC. *Approaches to Social Research*. 3rd ed., New York, NY: Oxford University Press; 1999.
18. Cohen L, Manion L, Morrison K. *Research Methods in Education*. 7th ed., London, UK: Routledge; 2011.
19. Smith SM, Albaum GS. *Fundamentals of Marketing Research*. Thousand Oaks, CA: SAGE Publications; 2005.
20. Andrich D. *Rasch Models for Measurement*. Newbury Park, CA: SAGE Publications; 1988.
21. Linacre BD, Wright JM. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil*. 1989;70(12):857–860.
22. Wolfe EW, Smith EV. Instrument development tools and activities for measure validation using Rasch models: part I—instrument development tools. *J Appl Meas*. 2007;8(1):97–123.
23. Norman G. Likert scales, levels of measurement and the "laws" of statistics. *Adv Health Sci Educ Theory Pract*. 2010;15(5):625–632.
24. Likert R. A technique for the measurement of attitudes. *Arch Psychol*. 1932;22(140):5–55.
25. Burns N, Grove SK. *The Practice of Nursing Research: Appraisal, Synthesis, and Generation of Evidence*. 6th ed., St. Louis, MO: Saunders; 2009.
26. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 5th ed., New York, NY: Oxford University Press; 2015.

27. Desselle SP. Construction, implementation, and analysis of summated rating scales. *Am J Pharm Educ*. 2005;69(5): Article 97.
28. Uebersax JS. Likert scales: dispelling the confusion. Statistical Methods for Rater Agreement website. 2006. Available at: ⟨http://john-uebersax.com/stat/likert.htm⟩ Accessed August 3, 2015.
29. Gaito J. Measurement scales and statistics: resurgence of an old misconception. *Psychol Bull*. 1980;87(3):564–567.
30. Lord FM. On the statistical treatment of football numbers. *Am Psychol*. 1953;8(12):750–751.
31. Wright DB. Football standings and measurement levels. *J R Stat Soc Ser D Statistician*. 1997;46(1):105–110.
32. Townsend JT, Ashby FG. Measurement scales and statistics: the misconception misconceived. *Psychol Bull*. 1984;96(2): 394–401.
33. Baker BO, Hardyck CD, Petrinovich LF. Weak measurements vs. strong statistics: an empirical critique of S. S. Stevens' proscriptions on statistics. *Educ Psychol Meas*. 1966;26(2): 291–309.
34. Labovitz S. Some observations on measurement and statistics. *Soc Forces*. 1967;46(2):151–160.
35. Labovitz S. The assignment of numbers to rank order categories. *Am Sociol Rev*. 1970;35(3):515–524.
36. Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev Educ Res*. 1972;42(3):237–288.
37. Baggaley AR, Hull AL. The effect of nonlinear transformations on a Likert scale. *Eval Health Prof*. 1983;6(4):483–491.
38. Dehaene S, Dupoux E, Mehler J. Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *J Exp Psychol Hum Percept Perform*. 1990;16(3):626–641.
39. Dehaene S, Bossini S, Giraux P. The mental representation of parity and number magnitude. *J Exp Psychol Gen*. 1993;122 (3):371–396.
40. Zorzi M, Pritis K, Umiltà C. Neglect disrupts the mental number line. *Nature*. 2002;417(6885):138–139.
41. Cohen DJ, Blanc-Goldhammer D. Numerical bias in bounded and unbounded number line tasks. *Psychon Bull Rev*. 2011;18 (2):331–338.
42. Fias W, Brysbaert M, Geypens F, dYdewalle G. The importance of magnitude information in numerical processing: evidence from the SNARC effect. *Math Cognit*. 1996;2(1): 95–110.
43. Gevers W, Verguts T, Reynvoet B, Caessens B, Fias W. Numbers and space: a computational model of the SNARC effect. *J Exp Psychol Hum Percept Perform*. 2006;32(1): 32–44.
44. Harvey BM, Klein BP, Petridou N, Dumoulin SO. Topographic representation of numerosity in the human parietal cortex. *Science*. 2013;341(6150):1123–1126.
45. Miller DC, Salkind NJ. *Handbook of Research Design and Social Measurement*. 6th ed., Thousand Oaks, CA: SAGE Publications; 2002.
46. DeVellis RF. *Scale Development: Theory and Applications*. 3rd ed., Thousand Oaks, CA: SAGE Publications; 2003.
47. Furr RM. *Scale Construction and Psychometrics for Social and Personality Psychology*. Thousand Oaks, CA: SAGE Publications; 2011.
48. Morgenstern H. Ecologic studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed., Philadelphia, PA: Lippincott Williams & Wilkins; 2008:511–531.
49. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs. Holm methods. *Am J Public Health*. 1996;86(5):726–728.
50. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43–46.
51. Greenland S, Rothman KJ. Fundamentals of epidemiologic data analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed., Philadelphia, PA: Lippincott Williams & Wilkins; 2008:213–237.
52. Perneger TV. What's wrong with Bonferroni adjustments? *Br Med J*. 1998;316(7139):1236–1238.
53. Robbins NM, Heiberger RM. *Plotting Likert and other rating scales. JSM Proceedings. Survey Research Methods Section*. Alexandria, VA: American Statistical Association; 2011, p. 1058–1066.
54. Johnson B, Christensen L. *Educational Research: Quantitative, Qualitative, and Mixed Approaches*. 3rd ed., Thousand Oaks, CA: SAGE Publications; 2008.
55. Fayers P. Alphas, betas, and skewy distributions: two ways of getting the wrong answer. *Adv Health Sci Educ Theory Pract*. 2011;16(3):291–296.
56. Schwarz N, Knäuper B, Hippler JH, NoelleNeumann E, Clark L. Numeric values may change the meaning of scale labels. *Public Opin Q*. 1991;55(4):570–582.
57. Schwarz N, Grayson CE, Knäuper B. Formal features of rating scales and the interpretation of question meaning. *Int J Public Opin Res*. 1998;10(2):177–183.
58. Friedman HH, Amoo T. Rating the rating scales. *J Mark Manag*. 1999;9(3):114–123.
59. Hsu TC, Feldt LS. The effect of limitations on the number of criterion score values on the significance level of the F-test. *Am Educ Res J*. 1969;6(4):515–527.
60. Johnson SM, Smith P, Tucker S. Response format of the job descriptive index: assessment of reliability and validity by the multitrait-multimethod matrix. *J Appl Psychol*. 1982;67(4):500–505.
61. Harpe SE, Phipps LB, Alowayesh MS. Effects of a learning-centered approach to assessment on students' attitudes towards and knowledge of statistics. *Curr Pharm Teach Learn*. 2012;4 (4):247–255.
62. Finney SJ, Schraw G. Self-efficacy beliefs in college statistics courses. *Contemp Educ Psychol*. 2003;28(2):161–186.
63. Schau C, Stevens J, Dauphinee TL, Del Vecchio A. The development and validation of the survey of attitudes toward statistics. *Educ Psychol Meas*. 1995;55(5):868–875.
64. Ferrill MJ, Norton LL, Blalock SJ. Determining the statistical knowledge of pharmacy practitioners: a survey and review of the literature. *Am J Pharm Educ*. 1999;63(4):371–376.
65. Harpe SE, Phipps LB. Evaluating student perceptions of a learner-centered drug literature evaluation course. *Am J Pharm Educ*. 2008;72(6): Article 135.
66. Huynh D, Haines ST, Plaza CM, et al. The impact of advanced pharmacy practice experiences on students' readiness for self-directed learning. *Am J Pharm Educ*. 2009;73 (4): Article 65.
67. Fisher M, King J, Tague G. Development of a self-directed learning readiness scale for nursing education. *Nurse Educ Today*. 2001;21(7):516–525.
68. Worley MM. Survey design. In: Aparasu RR, ed. *Research Methods for Pharmaceutical Practice and Policy*. London. UK: Pharmaceutical Press; 2011:161–177.

69. Hoover MJ, Jung R, Jacobs DM, Peeters MJ. Educational testing validity and reliability in pharmacy and medical education literature. *Am J Pharm Educ*. 2013;77(10): Article 213.

70. Peeters MJ, Beltyukova SA, Martin BA. Educational testing and validity of conclusions in the scholarship of teaching and learning. *Am J Pharm Educ*. 2013;77(9): Article 186.

71. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Appl Psychol Meas*. 1977;1 (3):385–401.

72. Ware JE. Jr, Sherbourne CD. The MOS 36-item short-form survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30(6):473–483.