Analyzing and Interpreting Data From Likert-Type Scales

Gail M. Sullivan, MD, MPH
Anthony R. Artino Jr, PhD

Likert-type scales are frequently used in medical education and medical education research. Common uses include end-of-rotation trainee feedback, faculty evaluations of trainees, and assessment of performance after an educational intervention. A sizable percentage of the educational research manuscripts submitted to the *Journal of Graduate Medical Education* employ a Likert scale for part or all of the outcome assessments. Thus, understanding the interpretation and analysis of data derived from Likert scales is imperative for those working in medical education and education research. The goal of this article is to provide readers who do not have extensive statistics background with the basics needed to understand these concepts.

Developed in 1932 by Rensis Likert[1] to measure attitudes, the typical Likert scale is a 5- or 7-point ordinal scale used by respondents to rate the degree to which they agree or disagree with a statement (TABLE). In an ordinal scale, responses can be rated or ranked, but the distance between responses is not measurable. Thus, the differences between "always," "often," and "sometimes" on a frequency response Likert scale are not necessarily equal. In other words, one cannot assume that the difference between responses is equidistant even though the numbers assigned to those responses are. This is in contrast to interval data, in which the difference between responses can be calculated and the numbers do refer to a measureable "something." An example of interval data would be numbers of procedures done per resident: a score of 3 means the resident has conducted 3 procedures. Interestingly, with computer technology, survey designers can create continuous measure scales that do provide interval responses as an alternative to a Likert scale. The various continuous measures for pain are well-known examples of this (FIGURE 1).

The Controversy

In the medical education literature, there has been a longstanding controversy regarding whether ordinal data, converted to numbers, can be treated as interval data.[2] That is, can means, standard deviations, and parametric statistics, which depend upon data that are normally distributed (FIGURE 2), be used to analyze ordinal data?

When conducting research, we measure data from a sample of the total population of interest, not from all members of the population. Parametric tests make assumptions about the underlying population from which the research data have been obtained—usually that these population data are normally distributed. Nonparametric tests do not make this assumption about the "shape" of the population from which the study data have been drawn. Nonparametric tests are less powerful than parametric tests and usually require a larger sample size (n value) to have the same power as parametric tests to find a difference between groups when a difference actually exists. Descriptive statistics, such as means and standard deviations, have unclear meanings when applied to Likert scale responses. For example, what does the average of "never" and "rarely" really mean? Does "rarely and a half" have a useful meaning?[3] Furthermore, if responses are clustered at the high and low extremes, the mean may appear to be the neutral or middle response, but this may not fairly characterize the data. This clustering of extremes is common, for example, in trainee evaluations of experiences that may be very popular with one group and perceived as unnecessary by others (eg, an epidemiology course in medical school). Other non-normal distributions of response data can similarly result in a mean score that is not a helpful measure of the data's central tendency.

Because of these observations, experts over the years have argued that the median should be used as the measure of central tendency for Likert scale data.[3] Similarly, experts have contended that frequencies (percentages of responses in each category), contingency tables, χ^2 tests, the

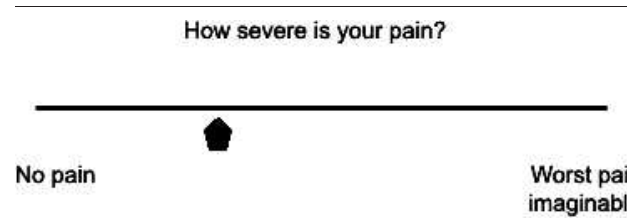

FIGURE 1 **Continuous Measure Example**

Please tell us your current pain level by sliding the pointer to the appropriate point along the continuous pain scale above.

Gail M. Sullivan, MD, MPH, is Editor-in-Chief of the *Journal of Graduate Medical Education*, and **Anthony R. Artino Jr, PhD,** is Associate Professor of Medicine and Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences.

Corresponding author: Gail M. Sullivan, MD, MPH, University of Connecticut, 253 Farmington Avenue, Farmington, CT 06030-5215, gsullivan@nso1.uchc.edu

DOI: http://dx.doi.org/10.4300/JGME-5-4-18

TABLE **Typical Likert Scales**

1	2	3	4	5
Never	Rarely	Sometimes	Often	Always
Completely disagree	Disagree	Neutral	Agree	Completely agree

Spearman rho assessment, or the Mann-Whitney *U* test should be used for analysis instead of parametric tests, which, strictly speaking, require interval data (eg, *t* tests, analysis of variance, Pearson correlations, regression).[3] However, other experts assert that if there is an adequate sample size (at least 5–10 observations per group) and if the data are normally distributed (or nearly normal), parametric tests can be used with Likert scale ordinal data.[3]

Fortunately, Dr. Geoff Norman, one of world's leaders in medical education research methodology, has comprehensively reviewed this controversy. He provides compelling evidence, with actual examples using real and simulated data, that parametric tests not only can be used with ordinal data, such as data from Likert scales, but also that parametric tests are generally more robust than nonparametric tests. That is, parametric tests tend to give "the right answer" even when statistical assumptions—such as a normal distribution of data—are violated, even to an extreme degree.[4] Thus, parametric tests are sufficiently robust to yield largely unbiased answers that are acceptably close to "the truth" when analyzing Likert scale responses.[4]

Educators and researchers also commonly create several Likert-type items, group them into a "survey scale," and then calculate a total score or mean score for the scale items. Often this practice is recommended, particularly when researchers are attempting to measure less concrete concepts, such as trainee motivation, patient satisfaction, and physician confidence—where a single survey item is unlikely to be capable of fully capturing the concept being assessed.[5] In these cases, experts suggest using the Cronbach alpha or Kappa test or factor analysis technique to provide evidence that the components of the scale are sufficiently intercorrelated and that the grouped items measure the underlying variable.

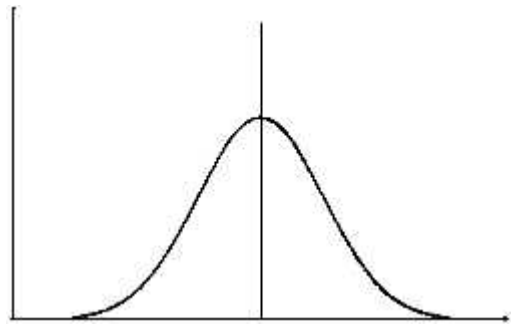

FIGURE 2 **A Normal Distribution**

The Bottom Line

Now that many experts have weighed in on this debate, the conclusions are fairly clear: parametric tests can be used to analyze Likert scale responses. However, to describe the data, means are often of limited value unless the data follow a classic normal distribution and a frequency distribution of responses will likely be more helpful. Furthermore, because the numbers derived from Likert scales represent ordinal responses, presentation of a mean to the 100th decimal place is usually not helpful or enlightening to readers.

In summary, we recommend that authors determine how they will describe and analyze their data as a first step in planning educational or research projects. Then they should discuss, in the Methods section or in a cover letter if the explanation is too lengthy, why they have chosen to portray and analyze their data in a particular way. Reviewers, readers, and especially editors will greatly appreciate this additional effort.

References

1 Likert R. A technique for the measurement of attitudes. *Arch Psychology*. 1932;22(140):55.
2 Carifio L, Perla R. Resolving the 50-year debate around using and misusing Likert scales. *Med Educ*. 2008;42(12):1150–1152.
3 Jamieson S. Likert scales: how to (ab)use them. *Med Educ*. 2004;38(12):1217–1218.
4 Norman G. Likert scales, levels of measurement and the "laws" of statistics. *Adv Health Sci Educ Theory Pract*. 2010;15(5):625–632.
5 Rickards G, Magee C, Artino AR Jr. You can't fix by analysis what you've spoiled by design: developing survey instruments and collecting validity evidence. *J Grad Med Educ*. 2012;4(4):407–410.