Reference

1. Baker K: Determining resident clinical performance: Getting beyond the noise. Anesthesiology 2011; 115:862–78

(Accepted for publication January 26, 2012.)

Likert or Not, We Are Biased

To the Editor:

I read with interest the recent article by Baker regarding the value of normalizing resident evaluation scores to eliminate individual faculty evaluator bias.[1] Without unduly undermining the importance of this study, I have concern about the statistical handling of Likert scores. Likert scores were used to create individual faculty member mean scores, faculty score standard deviations, and average resident scores when more than one core competency section was included. The central issue is that Likert scales involve ordinal data, or categories falling in a hierarchy.[2] Because the numbers in a Likert scale represent verbal statements of rank order (*e.g.*, 5 = distinctly above peer level), summarizing such ordinal data with a mean value is inappropriate by strict statistical methodology.[2] Moreover, the intervals between data points on a Likert scale are not necessarily equal or even certain.[3] To put this in the context of the study, consider this example from the relative performance designation used in the study: a score of "4" is "somewhat above peer level" and a score of "5" is "distinctly above peer level"; however, an average score of "4.5" cannot be said to represent "somewhat-above-peer-level-and-a-half."[4] Similarly, on the absolute/anchored competency designation, the difference between a score of "5" (performed in a fully independent manner) and a score of "6" (able to serve as a consultant to other physicians) is not necessarily equivalent to the difference between a score of "2" (needed moderate assistance) and a score of "3" (needed only minimal assistance). It is difficult to determine what, if any, limitation was imposed on the study as a result of this violation of statistical propriety. Nevertheless, although a purist may pine for cleaner data and analysis, this distraction can be mitigated by considering what Stevens wrote in 1946: "for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results."[5]

I look forward to future contributions from Baker. When I was a fellow his efforts sparked my interest in resident education and continue to do so now.

Nicholas C. Watson, M.D., UMass Memorial Medical Center, University of Massachusetts Medical School, Worcester, Massachusetts. nicholas.watson1@gmail.com

References

1. Baker K: Determining resident clinical performance: Getting beyond the noise. Anesthesiology 2011; 115:862–78
2. Triola MF: Elementary statistics, 5th ed. Reading, MA: Addison-Wesley Publishing Company, Inc., 1992, pp 16–8
3. Jamieson S: Likert scales: How to (ab)use them. Med Educ 2004; 38:1217–8
4. Kuzon WM Jr, Urbanchek MG, McCabe S: The seven deadly sins of statistical analysis. Ann Plast Surg 1996; 37:265–72
5. Stevens SS: On the theory of scales of measurement. Science 1946; 103:677–80

(Accepted for publication January 26, 2012.)

Errors in Assessment of Resident Performance

To the Editor:

In a recent innovative study, Baker used relative Z scores (Z_{rel}) to correct for observer bias in the assessment of 108 anesthesiology residents.[1] We have concerns about the statistical methodology used in this study and believe there is a need for caution before his approach is widely adopted.

Baker distinguishes three groups: those "reliably above average," "reliably below average," and "not reliably different from average." His criterion for identifying a resident who is above average is that 1.96 times the SEM for that individual's Z score (a 95% CI for the SEM) does not overlap with zero. A similar criterion is used to identify "below average" residents. This approach is problematic.

Although Baker identifies 30% of residents as "reliably below average," with sufficient assessments, 50% would be "reliably below average" because the width of the CIs would decrease. It is trivially true that, as long as the distribution is symmetric, 50% of people are "below average," but this does not imply that all "below average" residents require what Baker terms "performance interventions." Baker's Z scores could be applied to any group of residents, even a sample of entirely competent anesthesiologists, and would still identify a proportion as "below average." Without a clinically relevant benchmark, Baker's approach cannot be used to identify anesthetic competence.

In translating an overall assessment of 'anesthetic competence' into a Z score, Baker makes certain assumptions. One of these is that the competence of anesthesiologists is an underlying, continuous variable that can be normalized. Although this assumption cannot be validated, it can be simulated using a Monte-Carlo approach. Figure 1 shows the results of a single run of such a simulation. The assumptions are: that each of 100 individuals has intraindividual variation in Z_{rel} scores that is normally distributed, and that the mean score for each individual is offset by a value that is similarly, randomly sampled from a normal distribution ("interindividual variation"), with a known SD (SD_{adj}). As both the generated SD

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org).