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 underestimating 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 (eg, 5 = distinctly above peer level), summarizing such ordinal data with a mean value is inappropriate by strict statistical methodology. 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

Errors in Assessment of Resident Performance

To the Editor:
In a recent innovative study, Baker used relative Z scores (Zrel) to correct for observer bias in the assessment of 108 anesthesia 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 Zrel 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 (SDadj). As both the generated SD

References

Copyright © by the American Society of Anesthesiologists. Unauthorized reproduction of this article is prohibited.
Fig. 1. Simulated ranking of residents' Z scores. A representative run from a Monte Carlo simulation in which Baker's criteria are applied to a known distribution. Error bars are 95% CIs on the mean of each resident's Z_{rel} score. Thirty-three residents are classified as either "above" or "below" average (false positives, F-ve, shown in red), despite being within the underlying, generated "competence threshold" of ±1 SD (SD_{adj}). In this run, 31 residents are adjudged above average and 38 below average; there are no false negatives (F-ve).

for each individual's variation in competence, and the interindividual offset of his/her mean are known, Baker's approach can be tested against this standard. This simulation produces results that are similar to Baker's figure 5.

With this simulation, a "competence threshold" can be set, beyond which individuals are known to be outliers. Although it would be more usual to set the threshold at ±1.96 SD_{adj} (i.e., to assume that just 2.5% of individuals are performing "over" or "under"), the number of individuals Baker categorizes as above or below average (27% and 30%, respectively) suggests a threshold of about ±1 SD_{adj} which would on average identify 33.6% of residents as either above or below a threshold.

Figure 1 shows that with a competence threshold of just 1 SD_{adj} Baker's approach misclassifies 33 of the 100 individuals in this run as "reliably" above or below average, despite their underlying competence being within 1 SD of the mean competence. Use of a higher threshold would result in even more misclassification. The paradox of Baker's approach is that the greater the number of evaluations of each individual, the more likely he/she is to be misclassified. Running the current simulation 10,000 times shows that, on average, almost 36% of residents would be misclassified at a competence threshold of 1 SD_{adj}.

The annotated source code of our model (written in R, version 2.10.1; R Foundation for Statistical Computing, Vienna, Austria) is available in Supplemental Digital Content 1, http://links.lww.com/ALN/A838. The model is robust with repeated testing, and with moderate alterations to the SDs used; more extreme changes produce plots that are incompatible with Baker's results. We find it difficult to retain Baker's Z-score normalization without the conclusion that his subsequent interpretation is flawed. Misclassification of residents based on Baker's approach clearly has implications for their management and even their future careers.

A more robust analytic approach would be to use analysis of means, provided the assumption of an underlying, continuous, and normalizable "competence" can be justified, and clinical benchmarking can be established.

The authors thank Associate Professor Ross Ihaka, Ph.D., Department of Statistics, University of Auckland, Auckland, New Zealand, for checking the R code used, and Martin J. Turner, Ph.D., University of Sydney, Sydney, Australia, for comments on the manuscript.


References

In Reply:
I wish to thank the letter-writers for effectively articulating concerns with the paper.¹ Cattano essentially poses a question and a concern. He asks for an explanation of the progressive bias that occurs when faculty members evaluate CA1 to CA3 residents. I offer no certain mechanism for this finding. It was shown in the paper, in figure 7, that more senior residents garner higher degrees of faculty confidence in their skills as they progress in residency. Perhaps this causes a halo effect in normative rating even though the cohort is constant. One mechanism that can largely be ruled out is attrition. At most, we lose one resident per year because of attrition. Cattano raises the concern that faculty members will positively bias their evaluations of residents with the hope that residents will reciprocate and evaluate their teaching skills in a similarly positive manner. The concern that faculty members will assign positively biased clinical performance scores in an attempt to win higher teaching scores from the residents is a significant concern. We attempt to minimize this occurrence by distributing only anonymous evaluations to residents, and thus residents are unaware of which faculty members evaluated them. Similarly, faculty members are provided with anonymous resident evaluations of their teaching, and faculty members are unaware of which residents evaluated them. As mentioned in the paper, we no longer provide residents with the actual scores of their performance. Thus, residents don't know the evaluative numbers assigned by faculty members. Lastly, when individual faculty member teaching scores are regressed against the clinical performance scores, no important relationship is discernible between those same individuals and residents (data not shown).

Anesthesiology 2012; 116:1152–66

Copyright © by the American Society of Anesthesiologists. Unauthorized reproduction of this article is prohibited.