



Clinical supervision: what does it mean to be better?

Timothy Ramsay, PhD 

Received: 29 June 2017/Revised: 30 July 2017/Accepted: 28 August 2017/Published online: 1 September 2017
© Canadian Anesthesiologists' Society 2017

To the Editor,

I read with interest the paper entitled “Measurement of faculty anesthesiologists’ quality of clinical supervision has greater reliability when controlling for the lenience of the rating anesthesia resident: a retrospective cohort study” that recently appeared in the *Journal*.¹ I would like to point out what I believe to be a critical logical flaw in its conclusions.

The authors compare a new model for determining the quality of faculty supervision with an old model, concluding that the new model results in “greater detection of anesthesiologists with significantly better (or worse) clinical supervision scores.” The problem is that the two models are asking fundamentally different questions—the original compares averages of scores, whereas the new one compares the probability of obtaining a perfect score—and are therefore not comparable. Furthermore, as I illustrate with a simple example, the new model’s definition of quality is oversimplified and could lead to perverse conclusions. The new model is simply not a good substitute for the old one.

Faculty supervision is measured by the Oliveira Filho supervision scale,¹ which asks trainees to fill out a nine-item questionnaire rating a single experience of working with the supervisor in an operating room. Each item is ranked on a scale of 1–4, with 4 reflecting the best performance. The overall score is the average of the nine items. Thus, an individual item’s score can be as high as

4—if the supervisor scores perfectly on all items—or as low as 1. The original model compared supervisors on the basis of their average score, reasoning that a higher average score implies better supervision. The second model compares supervisors on their probability of obtaining a perfect score, reasoning that supervisors who obtain perfect scores more often must be better. To see the problem with this method, consider the following two hypothetical supervisors, John and Mary.

Mary prides herself on her supervision and receives a perfect score 75% of the time. The remaining 25% of the time she receives a rating of 4 on eight of the nine questions and but has only a 3 on the remaining one, giving her an overall score of 3.89 $[(8 \times 4 + 3)/9]$. Overall, her average score is 3.97 $[(0.75 \times 4) + (0.25 \times 3.89)]$. John is a truly excellent supervisor when sober but, tragically, suffers from a crippling substance abuse problem. When he is doing well (80% of the time), John gets uniformly perfect scores of 4. Sadly, on the other 20% of days, John receives a score of only 1. Overall, John’s average score is now 3.4 $[0.8 \times 4 + 0.2 \times 1]$

According to the original model, Mary’s average score of 3.97 appropriately puts her well above John with his average of 3.4. In the model that Dexter *et al.* report, which should allow more precise detection of better clinical supervision, John’s 80% probability of getting a perfect score ranks him as a better supervisor than Mary with her probability of only 75%.

Another way to explain the deficiency of the second model is as follows. The first model uses all the information from a carefully validated performance measurement scale consisting of nine questions, each of which has four possible levels. With $9 \times 4 = 36$ possible values, this measurement scale could represent a full

This submission is accompanied by a reply. Please see Can J Anesth 2017; 64: this issue.

T. Ramsay, PhD (✉)
The Ottawa Hospital Research Institute, Ottawa, ON, Canada
e-mail: tramsay@ohri.ca

continuum of performance from excellent to very poor. The second model effectively replaces this sensitive, nuanced scale by a single question: “Was your staff supervisor’s performance perfect?” This quite literally equates Mary’s “almost perfect” (3.97) with John’s “terrible” (1) score and leads to John being judged a better supervisor than Mary.

Conflicts of interest None declared.

Editorial Responsibility This submission was handled by Dr. Hilary P. Grocott, Editor-in-Chief, *Canadian Journal of Anesthesia*.

Reference

1. Dexter F, Ledolter J, Hindman BJ. Measurement of faculty anesthesiologists’ quality of clinical supervision has greater reliability when controlling for the leniency of the rating anesthesia resident: a retrospective cohort study. *Can J Anesth* 2017; 64: 643-55.