Can J Anesth/J Can Anesth (2017) 64:643–655
DOI 10.1007/s12630-017-0866-4

REPORTS OF ORIGINAL INVESTIGATIONS

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

La mesure de la qualité de supervision clinique des anesthésiologistes facultaires est plus fiable lorsqu'on tient compte de l'indulgence du résident en anesthésie responsable de l'évaluation: une étude de cohorte rétrospective

Franklin Dexter, MD, PhD · Johannes Ledolter, PhD · Bradley J. Hindman, MD

Received: 12 July 2016 / Revised: 30 January 2017 / Accepted: 13 March 2017 / Published online: 27 March 2017

Abstract

Background *Our department monitors the quality of anesthesiologists' clinical supervision and provides each anesthesiologist with periodic feedback. We hypothesized that greater differentiation among anesthesiologists' supervision scores could be obtained by adjusting for leniency of the rating resident.*

Methods *From July 1, 2013 to December 31, 2015, our department has utilized the de Oliveira Filho unidimensional nine-item supervision scale to assess the quality of clinical supervision provided by faculty as rated by residents. We examined all 13,664 ratings of the 97 anesthesiologists (ratees) by the 65 residents (raters). Testing for internal consistency among answers to questions (large Cronbach's alpha > 0.90) was performed to rule out that one or two questions accounted for leniency. Mixed-effects logistic regression was used to compare ratees while controlling for rater leniency vs using Student t tests without rater leniency.*

Results *The mean supervision scale score was calculated for each combination of the 65 raters and nine questions. The Cronbach's alpha was very large (0.977). The mean score was calculated for each of the 3,421 observed combinations of resident and anesthesiologist. The logits of the percentage of scores equal to the maximum value of 4.00 were normally distributed (residents, $P = 0.24$; anesthesiologists, $P = 0.50$). There were 20/97 anesthesiologists identified as significant outliers (13 with below average supervision scores and seven with better than average) using the mixed-effects logistic regression with rater leniency entered as a fixed effect but not by Student's t test. In contrast, there were three of 97 anesthesiologists identified as outliers (all three above average) using Student's t tests but not by logistic regression with leniency. The 20 vs 3 was significant ($P < 0.001$).*

Conclusions *Use of logistic regression with leniency results in greater detection of anesthesiologists with significantly better (or worse) clinical supervision scores than use of Student's t tests (i.e., without adjustment for rater leniency).*

Résumé

Contexte *Notre département contrôle la qualité de supervision clinique des anesthésiologistes et donne des rétroactions périodiques à chaque anesthésiologiste. Nous avons émis l'hypothèse qu'une plus grande différentiation dans les scores de supervision des anesthésiologistes serait*

F. Dexter, MD, PhD (✉)
Division of Management Consulting, University of Iowa, 200 Hawkins Drive, 6-JCP, Iowa City, IA 52242, USA
e-mail: Franklin-Dexter@UIowa.edu
URL: http://www.FranklinDexter.net

F. Dexter, MD, PhD · B. J. Hindman, MD
Department of Anesthesia, University of Iowa, Iowa City, IA, USA

J. Ledolter, PhD
Department of Management Sciences, University of Iowa, Iowa City, IA, USA

obtenue en tenant compte de l'indulgence du résident évaluateur.

Méthode *Dès le 1er juillet 2013 et jusqu'au 31 décembre 2015, notre département s'est servi d'une échelle de supervision unidimensionnelle à neuf points, soit l'échelle de Oliveira Filho, afin d'évaluer la qualité de la supervision clinique offerte par les membres du département et telle que notée par les résidents. Nous avons passé en revue les 13 664 évaluations des 97 anesthésiologistes (les évalués) par les 65 résidents (les évaluateurs). Nous avons évalué la cohérence interne parmi les réponses aux questions (coefficient alpha de Cronbach étendu > 0,90) afin d'écarter la possibilité qu'une ou deux questions expliqueraient l'indulgence. Une régression logistique à effets mixtes a été utilisée pour comparer les évalués tout en contrôlant l'indulgence des évaluateurs vs l'utilisation de tests t de Student sans indulgence de l'évaluateur.*

Résultats *Le score moyen sur l'échelle de supervision a été calculé pour chaque combinaison des 65 évaluateurs et des neuf questions. Le coefficient alpha de Cronbach était très étendu (0,977). La note moyenne a été calculée pour chacune des 3421 combinaisons observées de résident et anesthésiologiste. La distribution des logits du pourcentage de notes égales à la valeur maximale de 4,00 était normale (résidents, P = 0,24; anesthésiologistes, P = 0,50). Au total, 20/97 anesthésiologistes ont été identifiés comme valeurs aberrantes (13 ayant des notes de supervision au-dessous de la moyenne et sept des notes au-dessus) à l'aide de la régression logistique à effets mixtes avec l'indulgence de l'évaluateur en tant qu'effet fixe, mais pas par le test t de Student. Par contre, trois des 97 anesthésiologistes ont été identifiés comme valeurs aberrantes (les trois au-dessus de la moyenne) à l'aide des tests t de Student, mais pas par régression logistique avec l'indulgence. Les 20 vs 3 étaient significatifs (P < 0,001).*

Conclusion *L'utilisation de la régression logistique avec l'indulgence permet une plus grande détection des anesthésiologistes présentant des notes significativement meilleures (ou moins bonnes) de supervision clinique que l'utilisation de tests t de Student (c.-à-d. sans ajustement pour tenir compte de l'indulgence de l'évaluateur).*

Anesthesiologists' clinical competency will be assessed from completion of residency through to retirement. There are several limitations with using simulation for these assessments. Mannequin-based and/or multidisciplinary simulations can require expensive travel to a test centre and time away from clinical practice for both ratees and raters. Simulations often focus on management of a few crises of brief duration. In contrast, *in situ* assessments quantify anesthesiologists' clinical performance in the dynamic and unpredictable environment where they personally deliver care. This environment includes a range of large and unexpected problems, where anesthesiologists' roles include foreseeing and preventing problems and where social, team, and environmental factors influence anesthesiologists' effectiveness.[1,2] Thus, as part of an overall assessment of clinical competency, our department uses *in situ* assessments of individual anesthesiologists working in operating rooms and other procedural locations (henceforth referred to as "ORs") to determine how well they provide clinical supervision of anesthesia residents (Table 1).[3-14] Higher scores for clinical supervision are associated with fewer resident reports of errors with adverse effects on patients (Table 2.15)[11-13] and greater preference for the anesthesiologist to care for the rating resident's family (Table 2.7).[7]

Supervision, in this context, refers to all clinical oversight functions directed toward assuring the quality of clinical care whenever the anesthesiologist is not the sole anesthesia care provider.[3-5] The de Oliveira Filho unidimensional nine-item supervision instrument is a reliable scale used to assess the

Table 1 de Oliveira Filho *et al.'s* instrument[6] for measuring faculty anesthesiologists' supervision of residents during clinical operating room care

1. The faculty provided me timely, informal, nonthreatening comments on my performance and showed me ways to improve
2. The faculty was promptly available to help me solve problems with patients and procedures
3. The faculty used real clinical scenarios to stimulate my clinical reasoning, critical thinking, and theoretical learning
4. The faculty demonstrated theoretical knowledge, proficiency at procedures, ethical behaviour, and interest/compassion/respect for patients
5. The faculty was present during the critical moments of the anesthetic procedure (e.g., anesthesia induction, critical events, complications)
6. The faculty discussed with me the perianesthesia management of patients prior to starting an anesthetic procedure and accepted my suggestions, when appropriate
7. The faculty taught and demanded the implementation of safety measures during the perioperative period (e.g., anesthesia machine checkout, universal precautions, prevention of medication errors, etc.)
8. The faculty treated me respectfully and strived to create and maintain a pleasant environment during my clinical activities
9. The faculty gave me opportunities to perform procedures and encouraged my professional autonomy

All questions were presented daily in the same sequence. The wording differs from that developed,[6] only to the extent that a) the word "faculty" was used instead of "instructor" to be more closely aligned with the role of supervision of the nurse anesthetists, and b) the verb tense was changed to past tense because each evaluation was for a specific date working together

quality of supervision provided by each anesthesiologist (Table 1).[6-9] The scale measures all attributes of anesthesiologists' supervision of anesthesia residents (Table 2.1)[6,7,10-13] and has been shown, in multiple studies, to do this as a unidimensional construct (Table 2.2).[6,9,10,12] Low supervision scores are associated with written comments about the anesthesiologist being disrespectful, unprofessional, and/or teaching poorly that day (Table 2.18).[10,14,15] Scores increase when anesthesiologists receive individual feedback regarding the quality of their supervision (Table 2.17).[4] Scores are monitored daily and each anesthesiologist is provided with periodic feedback.[3,15]

The supervision scale's maximum value is 4.00, which corresponds to a response of 4 (i.e., "always") to each of the nine questions (Table 1).[6] Because of the ceiling effect, multiple scores of 4.00 reduce the scale's reliability[7,9,10,15] to differentiate performance among the anesthesiologists, even though such differentiation is mandatory (see Discussion).

We previously asked residents to provide a single evaluation for the overall quality of supervision they received from the department's faculty (i.e., as if intended as an evaluation of the residency program) (Table 2.14).[13] We compared those overall scores pairwise with the mean of each resident's evaluations of all individual anesthesiologists with whom they worked during the preceding eight months.[13] Both sets of scores showed considerable heterogeneity among the residents (e.g., some residents provided overall lower scores than those of other residents).[13] Consequently, our hypothesis was that greater differentiation among anesthesiologists' supervision scores could be obtained by incorporating scoring leniency by the resident (rater) into the statistical analysis (i.e., treating a high score as less meaningful when given by a resident who consistently provides high scores, in other words, lenient relative to other raters).[A]

Methods

The University of Iowa Institutional Review Board affirmed (June 8, 2016) that this investigation did not meet the regulatory definition of research in human subjects. Analyses were performed with de-identified data.

From July 1, 2013 to December 31, 2015, our department utilized the de Oliveira Filho supervision scale to assess the quality of clinical supervision by staff anesthesiologists (Table 1).[6,7] The cohort reported herein includes all rater evaluations of all staff anesthesiologists (ratees) over that 2.5-year period chosen for convenience. We used five six-month periods because we previously showed that six months was a sufficient duration in our department for nearly all ratees to receive evaluations and for an adequate number of unique raters to differentiate reliably among ratees using the supervision scale.[9,10,15]

The evaluation process consisted of daily, automated e-mail requests[16] to raters to evaluate the supervision provided by each ratee with whom they worked the previous day in an OR setting for at least one hour, including obstetrics and/or non-operating room anesthesia (e.g., radiation therapy).[4,8-10] Raters evaluated ratees' supervision by logging in to a secure webpage.[8] The raters could not submit their rating until each of the nine questions was answered with their choice of 1-4: 1 = never; 2 = rarely; 3 = frequently; or 4 = always (Table 1). The "score" for each evaluation was equal to the mean of the responses to the nine questions (Table 1). The scores remained confidential and were provided to the ratees periodically (every six months) only after averaging among multiple raters.[1,15,17]

Statistical analysis

If one or two of the nine questions resulted in leniency among raters, a potential intervention would have been to either modify the question(s) or provide an example of behaviour that should affect the answer to the question(s) (see Discussion). In contrast, if leniency were present throughout all questions, then an analysis of leniency would need to incorporate the average scores of the raters. The question whether leniency was present in a few *vs* all questions was addressed by analyzing the mean ratings for each combination of the 65 raters and nine questions. These means had sample sizes of at least 37 answers and mean sample sizes of 210 answers (i.e., sufficient to make the ranks of 1, 2, 3, or 4 into interval levels of measurement). Cronbach's alpha, a test for internal consistency among answers to questions,[B] was calculated using the resulting 65 × 9 matrix, equally weighting each rater. The confidence interval (CI) for Cronbach's alpha was calculated using the asymptotic method.[18]

[A] Leniency is the scientific term. We searched Google Scholar on December 8, 2016. There were 962 results from "rater leniency" OR "raters' leniency" OR "rating leniency" OR "leniency of the rater" OR "leniency of the raters". There were 93.4% fewer results for "rater heterogeneity" OR "raters' heterogeneity" OR "heterogeneity of the rater" OR "heterogeneity of the raters".

[B] See http://FDshort.com/CronbachSplitHalf, accessed February 2017. For each respondent, select four of the nine questions, calculate the mean score, and calculate the mean score of the other five questions. Calculate among all raters the correlation coefficient between the pairwise split-half mean scores. Repeat the process using all possible split halves of the nine questions. The mean of the correlation coefficients is Cronbach's alpha. This measure of internal consistency provides quantification for the reliability of the use of the score alone.

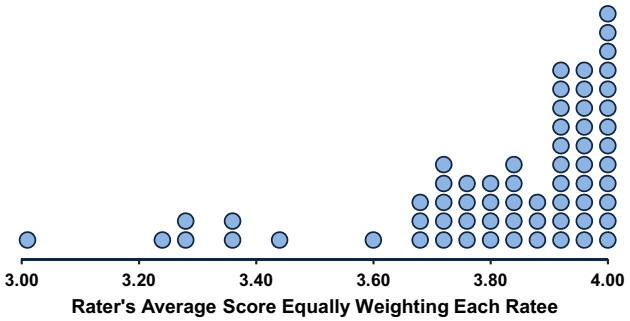

Fig. 1 Distribution of the average scores among raters. There were 3,421 observed combinations of the 65 raters (i.e., residents) and the 97 ratees (i.e., faculty anesthesiologists). For each combination, the mean was calculated. Then, for each rater, the average of the means across ratees was calculated. Those 65 values are shown in the figure. For details, see Footnote C and the companion papers of References (8) and (15). The dot plot shows that the distribution was unlike that of a normal distribution. The dot plot uses rounding for presentation; only one of the 65 raters had all scores equal to 4.00 (i.e., the average of the averages was also equal to 4.00)

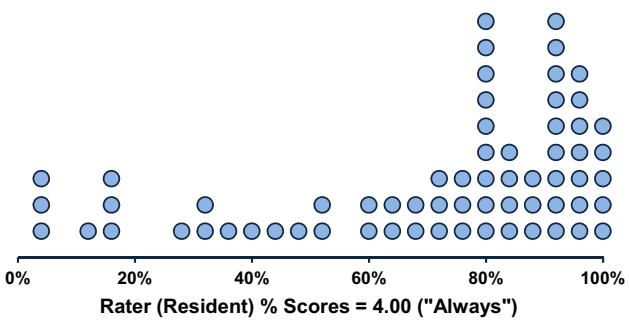

Fig. 2 Distribution among raters of their percentage of scores equal to the maximum value of 4.00 (i.e., all nine questions were answered "always"). The figure shows percentages for the 65 raters (i.e., residents). Unlike in Fig. 1, the 13,664 individual scores were used, and 9,305 (68.1%) were equal to 4.00. The dot plot shows substantial distribution of percentages among raters. The dot plot uses rounding for presentation; matching Fig. 1, there is just one rater with all scores equal to 4.00 (i.e., 100%). Among the 49 raters each with > 12 scores ≤ 4.00, the logits followed a normal distribution: Lilliefors test, $P = 0.24$. Among the 38 raters each with ≥ 19 such scores, Lilliefors test, $P = 0.21$

The statistical distribution of the rater leniency scores was analyzed. For each observed rater and ratee pair, we calculated the mean score specific to that pair.[8,15,19,C] For each rater, we obtained the equally weighted average of the mean scores provided by that rater for all ratees (Fig. 1).[8,C] We previously showed that the number of scores per pair differs markedly among raters for each ratee (i.e., there is non-random assignment of residents and anesthesiologists such that leniency will not average out; $P < 0.001$).[8,19]

The same approach was used when calculating the average of the means by ratee. In order to assess whether the scores of individual ratees were unusually low or high, we compared the averages of the means of each ratee with the value of 3.80 using Student's *t* test and Wilcoxon signed-rank test. The value of 3.80 is the overall mean supervision score among all ratees' scores (see Results). The *P* values using the Wilcoxon signed-rank test were exact, calculated using StatXact® 11 (Cytel Inc., Cambridge, MA, USA). Student's *t* test does not adjust for rater leniency.

Stata® 14.1 was used (StataCorp LP, College Station, TX, USA) to perform mixed-effects analyses treating the rater as a categorical fixed effect and the ratee as a random effect. Results of the mixed-effects analyses allowed us to assess the ratees' quality of supervision. Mixed-effects analyses were carried out for two separate dependent variables, modelled individually: 1) the average score, and 2) the binary variable whether the score equalled the maximum of 4.00 (Fig. 2). The logistic regression was performed using the "melogit" command option of mean-variance adaptive Gauss-Hermite quadrature with 30 quadrature points. As described later (see Results), several analyses were repeated using other estimation methods, including cluster-robust variance estimation, the unstructured covariance matrix, or estimation by Laplace approximation. All tests for differences between ratees were performed treating two-sided $P < 0.01$ as statistically significant. The imbalance in the number of resident ratings per ratee is considered in Appendix 1.

Results

Internal consistency of raters' answers to the nine questions contributing to the score

Individual questions did not contribute significantly to leniency (i.e., consideration of individual questions could not improve the statistical modelling). Cronbach's alpha as to the raters' answers to questions was high in value (0.977; 95% CI, 0.968 to 0.985); therefore, the score for each rating could be used (i.e., the mean of the answers to the nine questions in the supervision scale) (Table 1).

Statistical distributions of rater and ratee scores

We used the 13,664 scores,[D] with 3,421 observed combinations of the 65 raters and 97 ratees. In Appendix 2,

[C] The sample sizes are too small to estimate the variance within pairs, and the variances are generally unequal among pairs.[7,8] See the *Anesthesia & Analgesia* companion papers for mathematical details.[7,8] Even when there are many ratings per rater, using each rating's score minimally influences final assessments clinically.[19]

[D] Residents provided a response for 99.1% ($n = 14,585$) of the 14,722 requests.[10] For 6.3% ($n = 921$) of requests, residents responded that they worked with the faculty for insufficient time to evaluate supervision, leaving $n = 13,664$ ratings.[10] The mean (SD) intraoperative patient care time together was 4.87 (2.53) h day^{-1}.[10]

we show lack of validity of the statistical assumptions for a random effects model in the original score scale.[20,21]

We treated the rater as a fixed effect to incorporate rater leniency in a mixed-effects logistic regression model. Fig. 2 shows the distribution among raters of the percentage of scores equal to the maximum value of 4.00 (i.e., all nine questions answered "always"). The 65 raters differed significantly amongst each other in terms of the percentages of their scores equal to 4.00 ($P < 0.001$ using fixed-effect logistic regression).

The mixed-effects model with rater as a fixed effect and ratee as a random effect relies on the assumption that the distribution of the logits among ratees follows a normal distribution, which it does. Specifically, no ratee had all scores equal to 4.00 (i.e., for which the logit would have been undefined because of division by zero). In addition, no ratee had all scores less than 4.00 (i.e., for which the logit would also be undefined). There were 60 ratees each with $\geq$ 14 scores lower than 4.00 among their $\geq$ 32 scores (i.e., sample sizes large enough to obtain reliable estimates of the logits).[8,22] The logits followed a normal distribution [Lilliefors test, $P = 0.50$; mean (standard deviation [SD]), −0.781(0.491)].

Effectiveness of logistic regression with leniency relative to Student's t tests (i.e., without adjustment for leniency)

Figures 3 and 4 show each ratee's average score, meaning the average of means, equally weighting each rater (see above "Statistical distributions of rater and ratee scores").[4,8,10,15] We use symbols to indicate whether the ratee's average score is significantly ($P < 0.01$) different from the average score among all ratees when using a Student's *t* test (i.e., without adjustment for rater leniency). We also indicate whether the ratee's percentage of scores < 4.00 differed from other ratees when using mixed-effects logistic regression, with rater leniency treated as a fixed effect and ratee as a random effect. We subsequently refer to that mixed-effects model as "logistic regression with leniency".

The principal result is that 20/97 ratees were identified as outliers using the logistic regression with leniency, but not by Student's *t* tests. There were 3/97 ratees identified as outliers using the Student's *t* tests, but not by logistic regression with leniency. The 20 *vs* 3 is significant; exact $P < 0.001$ using McNemar's test. Thus, adjusting for rater leniency increased the ability to distinguish the quality of anesthesiologists' clinical supervision.

In Appendix 3, we confirm the corollary that there is less information from scores < 4.00 *vs* the percentage of scores equal to the maximum score of 4.00.

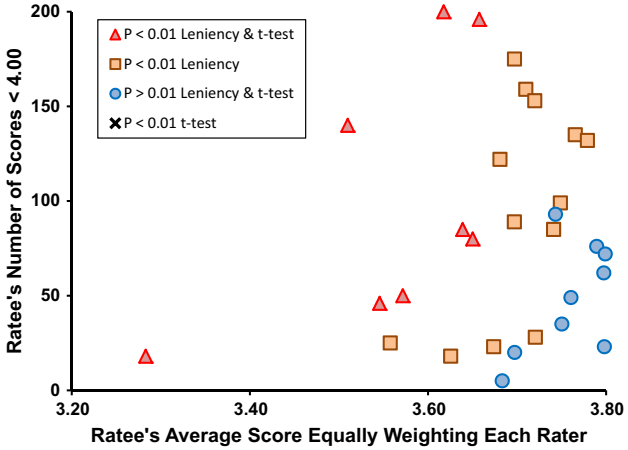

Fig. 3 Comparison between Student's *t* test and logistic regression with leniency for identification of ratees with below average quality of supervision. Among the 97 ratees (i.e., faculty anesthesiologists), there were 14 ratees with less than the minimum of nine different raters (i.e., residents) for reliability. Those 14 ratees were excluded. Among the other 83 ratees, there were 30 with average scores < 3.80, with the averages calculated as in Fig. 1, equally weighting each rater. The data for the 30 ratees are shown. Among the 30 ratees, there are eight plotted using red triangles. Student's *t* test found that the averages for the eight ratees were < 3.80 ($P < 0.01$). The mixed-effects logistic regression with rater leniency treated as a fixed effect also found that the percentage of scores equal to 4.00 for each of these eight ratees differed from that of other anesthesiologists ($P < 0.01$). There are 13 ratees plotted using orange squares. These ratees were not significant by the Student's *t* test ($P \geq 0.025$), but were significant by logistic regression with leniency ($P \leq 0.0068$). There are nine ratees plotted using blue circles. Neither of the two statistical methods found these ratees to be significant at $P < 0.01$. There were 0 (i.e., no) ratees for which the Student's *t* test was significant, but not so for logistic regression. Therefore, the black X is shown in the legend to match Fig. 5, but no such data points are plotted. The scale of the vertical axis in Fig. 3 differs from that in Fig. 4

In Appendix 4, we show that our previous observation of an increase in supervision score over time with evaluation and feedback (Table 2.17)[4] holds when analyzed using logistic regression with leniency.

In Appendix 5, we show that our previous analyses and publications without consideration of rater leniency were reasonable because initially there was greater heterogeneity of scores among ratees.

Graphical presentation of the principal result

In this final section, we examine why incorporating rater leniency increased the sensitivity to detect both below average and above average performance differences among ratees. Readers who are less interested in "why" may want to go directly to the Discussion.

The figures are divided between descriptions of the supervision scores of ratees (i.e., anesthesiologists) with

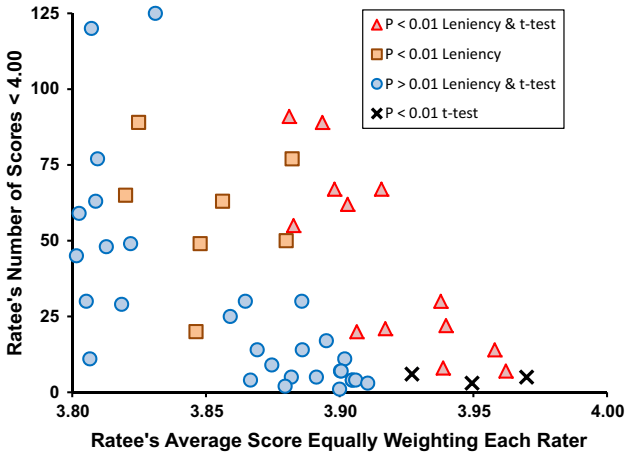

Fig. 4 Comparison between Student's *t* test and logistic regression with leniency for identification of ratees with greater than average quality of supervision. This figure matches Fig. 3, except that it is for the 53 ratees (i.e., faculty anesthesiologists) with average scores > 3.80. Among the 53 ratees, there are 13 plotted using red triangles. Student's *t* test found these 13 ratees' averages to be > 3.80 ($P < 0.01$). The mixed-effects logistic regression with rater leniency treated as a fixed effect also found that the percentage of scores equal to 4.00 for each of these 13 ratees differed from that of other anesthesiologists ($P < 0.01$). There are seven ratees plotted using orange squares. These ratees were not significant by the Student's *t* test ($P \geq 0.016$), but were significant by logistic regression with leniency ($P \leq 0.0034$). There are 30 ratees plotted using blue circles. Neither of the two statistical methods found these ratees to be significant at $P < 0.01$. There are three ratees plotted using black Xs. These were significant by the Student's *t* test ($P \leq 0.0018$) but not by logistic regression with leniency ($P \geq 0.32$). The scale of the vertical axis in Fig. 4 differs from that in Fig. 3

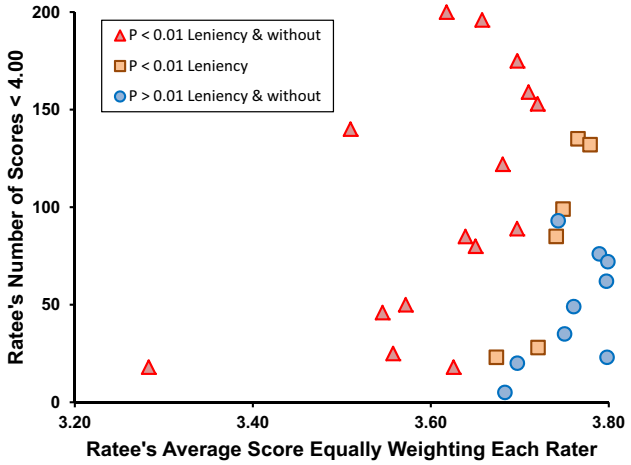

Fig. 5 Comparison between logistic regression with and without leniency for identification of ratees with below average quality of supervision. As described in Fig. 3 legend, data for 30 ratees (i.e., faculty anesthesiologists) are displayed, each with average scores < 3.80. Among the 30 ratees, 15 are plotted using red triangles. Both logistic regression models detected that these ratees had a significantly ($P < 0.01$) greater percentage of scores < 4.00 than other ratees. There are six ratees plotted using orange squares. Logistic regression with leniency, but not without leniency, detected these ratees as providing significantly lower quality supervision than other ratees. There are nine ratees plotted using blue circles. Neither of the two statistical methods found these ratees to be significant at $P < 0.01$. Including leniency in the logistic regression did not prevent significance for any (0) ratees

scores less than (Figs 3, 5) and greater than (Figs 4, 6) the overall average score of 3.80. No ratee happened to have an average score equal to 3.80. Thus, the division reflected different concerns, i.e., identification of ratees potentially performing below average *vs* those potentially performing above average.

Statistical significance of the logistic regression with leniency depended on the number of scores < 4.00, shown on the vertical axes of Figs 3-6 (see Appendix 1 and Appendix 6). For a given ratee average score, blue circles showing lack of statistical significance are more often present for smaller sample sizes than red triangles and orange squares.

Among the 30 ratees with average scores < 3.80, 13 were not significantly different from the average of 3.80 using the Student's *t* test, but were significantly different from the other ratees by logistic regression with leniency (Fig. 3). For illustration, we consider the ratee with an average score of 3.56, shown by the left-most orange square. This score was the smallest value not found to be significantly less than the overall average of 3.80 using the Student's *t* test, but found to differ significantly from the other ratees by logistic regression with leniency. In Appendix 7, we show that this finding was caused by substantial variability among raters (i.e., residents) regarding how much the ratee's quality of supervision was less than the maximum score (4.00).

Among the 53 ratees who had average supervision scores > 3.80 and who had at least nine different raters, seven were not significantly different from average as determined by the Student's *t* test, but were significantly different using logistic regression with leniency (Fig. 4). There were 3/53 ratees who were significantly different from average by the Student's *t* test, but not significantly different using logistic regression with leniency. For illustration, we consider the ratee with the highest average score. In Appendix 8, we show that logistic regression with, or without, leniency (Fig. 6) lacked statistical power to differentiate this ratee from other anesthesiologists because the ratee had above average quality of supervision and relatively few clinical days (i.e., ratings).

Discussion

The supervision scores are the cumulative result of how the anesthesiologists perform in clinical environments. The scores reflect *in situ* performance and can improve with feedback.[4,15] Supervision scores are used in our department for mandatory annual collegiate evaluations and for

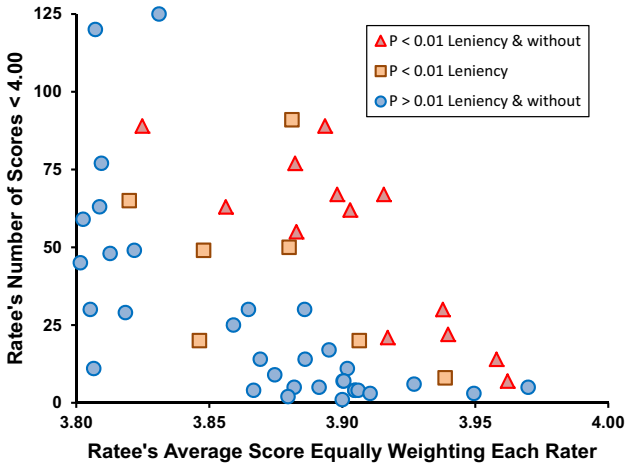

Fig. 6 Comparison between logistic regression with and without leniency for identification of ratees with greater than average quality of supervision. This figure matches Fig. 4, except that it is for the 53 ratees (i.e., faculty anesthesiologists) with average scores > 3.80. Among the 53 ratees, 13 are plotted using red triangles. Both logistic regression models detected that these ratees had a significantly ($P < 0.01$) greater percentage of scores < 4.00 than other ratees. There are seven ratees plotted using orange squares. Logistic regression with leniency, but not without leniency, detected these ratees as providing significantly lower quality supervision than other ratees. There are 30 ratees plotted using blue circles. Neither of the two statistical methods found these ratees to be significant at $P < 0.01$. Including leniency in the logistic regression did not prevent significance for any (0) ratees

maintenance of hospital clinical privileges (i.e., the United States' mandatory semi-annual "Ongoing Professional Practice Evaluation"). Consequently, the statistical comparisons could reasonably be considered to represent high-stakes testing.[E] We therefore considered statistical approaches that satisfy statistical assumptions as much as possible. In addition, we conservatively treated as statistically significant only those differences in ratee scores with small P values < 0.01 and used random effects modelling (i.e., shrinkage of estimates for anesthesiologists with small sample sizes toward the average).[23-27] Nevertheless, we show mixed-effects logistic regression modelling, with rater leniency entered as a fixed effect, which resulted in greater detection of performance outliers than with the Student's t test (i.e., without adjustment for rater leniency). Comparing the mixed-effects logistic regression model with rater leniency with multiple Student's t tests, rather than with a random effects model of the average scores without rater leniency, resulted in a lesser chance[23-25] of detecting benefit in logistic regression (i.e., our conclusion is deliberately conservative).

Previous psychometric studies of anesthesiologists' assessments of resident performance have also found significant rater leniency.[28,29] Even with an adjustment of the average scores for rater leniency, the number of different ratings that faculty needed for a reliable assessment of resident performance exceeded the total number of faculty in many departments.[28] Our paper provides a methodological framework for future statistical analyses of leniency for such applications.

Suppose the anesthesiologists were distributed into nine categories. There are those with a less than average, average, and greater than average annual number of clinical days, thereby receiving a less than average, average, and greater than average number of evaluations of their clinical performance. There are anesthesiologists who provide less than average, average, and greater than average quality of supervision. We think that, among these nine (3 × 3) groups, the least institutional cost for misclassifying the quality of clinical supervision (below average, average, above average) would be to consider the group of anesthesiologists providing less than average clinical workload and greater than average quality of supervision as providing average quality of supervision. Because this was the only group that was "misclassified" through use of logistic regression with leniency, we think it is reasonable managerially to use this method to analyze the supervision data.

We showed that leniency in the supervision scale (Table 1) was caused by the cumulative effect of all questions (i.e., leniency was not the disproportionate effect of a few questions). If an individual question had accounted for variability in leniency among raters, providing examples of behaviour corresponding to an answer could have been an alternative intervention to reduce leniency. Because our department provides OR care for a large diversity of procedures, it is not obvious to us how to provide examples because there are so many different interactions between residents and anesthesiologists that could contribute to less than or greater than average quality of supervision.[1,10] Nevertheless, the finding that leniency arises because of the cumulative effect of all questions shows that the issue is moot. Variability in rater leniency is the result of the raters' overall (omnibus) assessments of anesthesiologists' performance, without distinction among the nine items describing specific attributes of supervision.

The supervision score is a surrogate for whether a resident would choose the anesthesiologist to care for their family (Table 2.7).[7] Supervision scores for specific rotations are associated with perceived teamwork during the rotation (Table 2.8).[12] Observation of intraoperative briefings has found that sometimes anesthesiologists barely participate (e.g., being occupied with other activities).[30] Team members can "rate the value" of the intraoperative briefing performed "in the OR when the patient is awake".[31] Thus, we have hypothesized that leniency may

[E] High-stakes Testing. Wikipedia. Available from URL: https://en.wikipedia.org/wiki/High-stakes_testing (accessed February 2017).

be related to interaction among organizational safety culture, residents' perceptions of the importance of the intraoperative briefing to patient outcome, and the anesthesiologists' participation (or lack) in the briefings. Our finding of large internal rater consistency among the nine questions shows that such a hypothesis cannot be supported. Supervision begins when residents and anesthesiologists are assigned cases together, ends after the day's patient care is completed, and includes inseparable attributes (Table 1). Future studies could evaluate whether rater leniency is personality based and/or applies to rating other domains such as quality of life.

Our findings are limited by raters being nested within departments (i.e., residents in one department rarely work with anesthesiologists in other departments). Consequently, for external reporting, we recommend that evaluation of each ratee (anesthesiologist, subspecialty,[12] or department[11] be performed using the equally weighed average of the scores from each rater. Results are reported as average scores of equally weighted raters, along with confidence intervals.[8,C] In contrast, for assessment and progressive quality improvement within a department, we recommend the use of mixed-effects logistic regression with rater leniency. Results are reported as odds ratios, along with confidence intervals. Regardless, *in situ* assessment of the quality of supervision depends (Figs 4 and 6) on there being at least nine (and preferably more) unique raters for each ratee (Table 2.11).[7] Although this generally holds for operating room anesthesia, it can be a limitation for specialties (e.g., chronic pain) in which residents rotate for weeks at a time and work with one or two attending physicians.

Conflicts of interest None declared.

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

Author contributions *Franklin Dexter* and *Bradley J. Hindman* helped design the study. *Franklin Dexter* helped conduct the study. *Franklin Dexter* and *Johannes Ledolter* helped analyze the data. *Franklin Dexter, Johannes Ledolter,* and *Bradley J. Hindman* helped write the manuscript.

Funding Departmental funding

Appendix 1 Imbalance in numbers of raters among ratees

Figures 3-6 control graphically for heterogeneity among ratees in average scores (i.e., horizontal axes) and in sample sizes (i.e., vertical axes), but not for imbalance in the relative proportions of different raters. Among the 3,411 combinations of rater and ratee, the most common number of ratings was one occasion together (26.2%); whereas the median was three occasions and the 90th percentile was nine occasions. The imbalance results in substantial residual scatter in the figures, even though the imbalance is taken into account in our inferential analysis. When all residents evaluated all anesthesiologists during one weekend in a balanced design, there was a statistically significant but quantitatively negligible effect of resident on supervision scores (Table 2.6).[7] Addition of resident class (i.e., "CA-1," "CA-2," or "CA-3") as another variable in the logistic regression with leniency failed because the sample sizes were too small to estimate the fixed effects when stratified by six-month period. A better choice for the horizontal axis would be the marginal estimate of the average score (i.e., adjusting for the relative proportions of different raters among ratees by treating both rater and ratee as fixed effects and then re-estimating the mean using the mean proportions). Nevertheless, there is substantial collinearity, and estimates are not calculated by Stata®.

Appendix 2 Lack of validity of assumptions for random effects model in score scale

From among the 13,664 scores,[D] there were 3,421 observed combinations of the 65 raters (i.e., residents) and 97 ratees (i.e., anesthesiologists). The mean score was calculated for each of the 3,421 combinations. The average of these means across ratees was then calculated for each of the 65 raters (Fig. 1).

The distribution among the raters of their $n = 65$ averages was skewed ($P < 0.0001$; skewness [G1] −1.82, standard error 0.30).[20] The 65 averages had sample mean (SD) of 3.80 (0.22). The corresponding normal distribution had poor fit to the data: Lilliefors test, $P < 0.0001$. After each of six power transformations[21] ($X^{-2.0}$, $X^{-1.0}$, $X^{-0.5}$, ln (X), $X^{0.5}$, and $X^{2.0}$), still $P < 0.0001$, showing the $n = 65$ data by rater not following normal distribution. Thus, treating rater leniency as a random effect would not work reliably because we would need falsely to assume a normal distribution. These calculations were performed using SYSTAT 13.1 (SYSTAT Software Inc., San Jose, CA, USA).

Next, we considered the probability distribution for the ratee effect. Mixed-effects modelling of scores with rater treated as a fixed effect and ratee as a random effect makes normality assumptions about the random effect, which needs to be checked. The average of the means for each combination was calculated for each of the 97 ratees. The normal distribution had sample mean (SD) of 3.79 (0.17), but poor fit: Lilliefors test, $P < 0.0001$. Normal

distributions were not suitable models for ratee averages, even after considering each of the six power transformations (all $P < 0.0001$).

Appendix 3 Relative information from scores < 4.00 *vs* percentage of scores = 4.00

Data presented in the Results suggest less information from scores < 4.00 *vs* the percentage of scores equal to the maximum scores of 4.00. As described in the Section "Effectiveness of logistic regression with leniency relative to Student's *t* test", there were significantly more (20/97) ratees identified as outliers using the logistic regression with leniency but not by Student's *t* test ($P = 0.0005$ by McNemar's test). Also, in the Section "Influence of leniency on evaluation of ratees with below average supervision", we provide an example wherein the Student's *t* test failed to detect a significantly less than the overall average score because of large variability among raters regarding how much the ratees' quality of supervision was less than maximum (4.00).

The most common value of the scores was 4.00 (Table 2.5).[9] The incidence of 4.00 among the 13,664 scores was 68.1%, significantly greater than 50% ($P < 0.00001$ by binomial test, StatXact 11). Consequently, suppose that information from scores < 4.00 were providing information highly correlated with and supplementing the percentage of scores equal to 4.00. Then, the next most common score would be the next smaller possible score, 3.89; the 3.89 = (8 questions × score of 4 + 1 question × score of 3) / (9 questions). This was not so. The next most common score was 3.00, with 8.90% of scores, significantly greater than the 4.30% of scores accounted for by the value of 3.89 ($P < 0.001$ by Fisher's exact test). The reason for this finding is probably that each of the four possible answers for each question represents an associated frequency: 1 = never; 2 = rarely; 3 = frequently; or 4 = always. Less than "frequently" has been associated with "mistakes that had negative consequences for the patient" (Table 2.15),[11] "medication errors (dose or incorrect drug) in the last year" (Table 2.15),[11] and disrespectful behaviour of the ratee (Table 2.18; $P < 0.0001$).[10]

An anonymous reviewer recognized an alternative potential explanation for our findings. Suppose that ratee A has 60% of scores = 4.00, but among scores < 4.00, the mean equals 2.00. Suppose that ratee B has 50% of scores = 4.00, but among scores < 4.00, the mean equals 3.80. The logistic model would then suggest that ratee A provides greater quality of supervision than ratee B, while the Student *t* test would suggest the opposite. The example highlights that the percentage of scores equal to the maximum value of 4.00 could be negatively correlated with the mean of all the ratee's scores < 4.00. This was not so. Among the 89 ratees with at least nine scores < 4.00, the Kendall's $\tau_b = +0.36$ (two-sided $P < 0.0001$) (StatXact 11). Furthermore, in the fifth six-month period (i.e., after considerable duration of feedback), among the 58 ratees with at least nine scores, Kendall's $\tau_b = +0.024$ ($P = 0.79$ relative to 0.00 and $P = 0.0003$ relative to 0.36). Thus, as supervision scores increased progressively, there was significantly less information provided by the scores < 4.00.

Appendix 4 Change in supervision scores with evaluation and feedback

We reported previously that, with evaluation and feedback, there was an increase in anesthesiologists' supervision scores from the first to the third six-month period (Table 2.17).[4] We re-evaluated that finding[4] using mixed-effects logistic regression, with rater leniency entered as a fixed effect. The six-month period, as continuous variable, was entered as a fixed effect. Each increase of one period was associated with an odds ratio of 1.34 (95% confidence interval [CI], 1.25 to 1.44) for a score equal to 4.00. Repeating the analysis using robust variance estimation, 95% CI was 1.20 to 1.51, and repeating the analysis using an unstructured covariance matrix, 95% CI was 1.25 to 1.44. Each of these three methods of analyses had $P < 0.0001$. Therefore, our previous observation of an increase in supervision scores over time[4] held even when rater leniency was included in the analysis.

Appendix 5 Change over time in heterogeneity of scores among anesthesiologists

Although our department had evaluated anesthesiologists' supervision for 2.5 years (five sequential six-month periods), we had not considered rater leniency in our analysis.

The sample estimate of the standard deviation (SD) among ratees in their average of mean scores was calculated for each of the five six-month periods. During the first six-month period, 57 anesthesiologists received scores from at least nine raters (i.e., residents). The SD among the raters' averages was 0.154 ($n = 57$). Among the subsequent four six-month periods, the SDs were 0.150 ($n = 56$), 0.120 ($n = 50$), 0.095 ($n = 55$), and 0.097 ($n = 56$), respectively.

Bartlett's test and Levene's test were used to test for inequality in the SDs among these periods as a check on the

reliability of the results. That calculation was performed using SYSTAT 13.1.

The lesser variance among ratees was a significant change (Bartlett's test, $P = 0.0002$ and Levene's test, $P = 0.0002$), prompting our consideration of statistical methods better able to distinguish among anesthesiologists.

Appendix 6 Ratees not included in the figures

Even using two figures (Figs 3 and 4), there was considerable data point overlap. In our previous generalizability analysis, when all raters (i.e., residents) evaluated all ratees (i.e., anesthesiologists) on the same weekend in a balanced design, a minimum of nine different raters was needed for reliability (Table 2.2).[7] Among the 97 ratees, there were 14 ratees who received scores from less than nine different raters. None of these 14 ratees differed significantly from the other ratees when we used mixed-effects logistic regression with rater leniency treated as a fixed effect (all $P \geq 0.041$). Therefore, to reduce clutter in Figs 3 and 4 (and in the subsequent Fig 5 and 6), the results for these 14 ratees are not shown graphically.

Appendix 7 Illustrative example of the ratee shown by the left-most orange square in Fig. 3

We consider the ratee with an average score of 3.56, shown by the left-most orange square in Fig. 3. This score was the smallest value not found to be significantly less than the overall average of 3.80 by Student's t test but for which the ratee was found to differ significantly from the other ratees by logistic regression with leniency. To understand the result in the logit scale, we consider the simpler process of logistic regression without consideration of leniency; these findings are shown in Fig. 5. This ratee had 32.4% of $n = 37$ scores equal to 4.00, considerably less than the overall percentage of 68.1%. This is a significant difference with $P < 0.0001$, by logistic regression both with and without leniency. Thus, the relevant question is why this ratee's average supervision score of 3.56 was not significantly different from 3.80 by Student's t test. This ratee had a sample standard deviation of 0.53. The number of raters was 20. Among the 12 of 20 raters providing averages < 3.80, the median was 3.43, with averages ranging from 2.39 to 3.78. Among the eight raters providing averages > 3.80, the median was 4.00, with averages ranging from 3.89 to 4.00. The distribution of averages below and above 3.80 was not significant by Wilcoxon signed-rank test ($P = 0.17$). Thus, the lack of significance of the Student's t test for this ratee was not due to limitations of the assumption of a normal distribution. Rather, there was substantial variability among raters in how much the quality of supervision provided by the ratee was less than maximum (4.00).

For this illustration, logistic regression models both with and without leniency were equally effective for differentiating the ratee's quality of supervision from other ratees. This was convenient for understanding the results. Nevertheless, Fig. 5 shows that the example represented an unusual finding.

Appendix 8 Illustrative example of the ratee with the largest average score (Fig. 4)

We consider the ratee with the largest average score. The average of the means was equal to 3.97, the standard deviation was 0.068, and $n = 16$ raters providing 21 individual scores, giving $P < 0.0001$ by Student's t test. For this ratee, among the averages, 13 were equal to 4.00, and the other three were equal to 3.89, 3.85, and 3.78. This distribution of averages relative to 3.80 was significant by Wilcoxon signed-rank test ($P < 0.0001$). Thus, identifying this ratee as having greater than the overall quality of supervision was not artifactual, caused by assumptions of the statistical distribution of averages. Thus, the relevant question is why this ratee's logistic regression with leniency was not significant ($P = 0.32$). We rely on the results of logistic regression without leniency (Fig. 6) because that model was easier to interpret and was also not significant ($P = 0.57$). There were 16/21 (76.2%) individual scores equal to 4.00. That percentage (i.e., the 76.2%) differed little from the overall percentage of 68.1%. To have an 80% power to detect such a difference of a single percentage at $\alpha \leq 0.01$ would require a sample size of 327 scores (i.e., greater than tenfold larger). Thus, use of logistic regression with leniency lacked statistical power to differentiate this ratee with relatively few ratings (i.e., few clinical days) but above average quality of supervision compared with other anesthesiologists.

The same thing applied to logistic regression without leniency. The observed percentage of scores equal to 4.00 was 76.2%, an 8.1% difference from the overall 68.1%. Among the 41 of 97 ratees deemed significant ($P < 0.01$) using logistic regression with leniency, 11 had absolute differences less than 8.1%. One of the 11 was also significant by logistic regression without leniency ($P = 0.0025$), with 61.8% of the scores equal to 4.00. This was possible because they had 122 scores less than 4.00 (i.e., a much larger sample size). This ratee is viewable in Fig. 5 at an average score of 3.68. Otherwise, incorporating leniency mattered. For example, a ratee had 28 evaluations less than 4.00 and an average score of 3.72 (Fig. 5). The ratee had 67.1% of scores equal to 4.00, an absolute difference of just

1.0%. $P = 0.0012$ incorporating leniency *vs* $P = 0.64$ without. For example, a ratee had 20 evaluations less than 4.00 and an average score of 3.85 (Fig. 6). The ratee had 73.0% of scores equal to 4.00, an absolute difference of 4.9%. $P = 0.0034$ incorporating leniency *vs* $P = 0.54$ without. For example, a ratee had 20 evaluations less than 4.00 and an average score of 3.91 (Fig. 6). $P = 0.0004$ incorporating leniency and $P = 0.46$ without.

Table 2 Previous findings regarding supervision of anesthesia residents and nurse anesthetists by faculty anesthesiologists

1. Supervision is a single-dimensional construct that incorporates several different attributes, including participation in perianesthesia planning, availability for help/consultation, presence during critical phases of the anesthetic, and fostering safety measures[6,7,10,12]
2. Supervision can be quantified reliably using an instrument with nine questions, each question assessing a different attribute of supervision.[6] The nine questions take < 90 sec to complete. The Cronbach's alpha achieved in routine use was equal to 0.948 ± 0.001 (SE)[10]
3. Raters evaluate how often each attribute is demonstrated by the anesthesiologist (never = 1; rarely = 2; frequently = 3; and always = 4), and the supervision score is the mean of the nine answers.[6] When each anesthesiologist's mean resident and mean nurse anesthetist scores were paired, the means were correlated ($P < 0.0001$).[8,9] Thus, the behaviour and attributes used to assess the quality of an anesthesiologist's supervision have significant commonality between residents and nurse anesthetists[8,9]
4. There were very small differences in anesthesiologist supervision scores provided by residents when 1) a resident had more units of work that day with the rated anesthesiologist; ("units together", $\tau_b = 0.083 \pm 0.014$) or 2) the rated staff anesthesiologist had more units of work that same day with other providers, ("units not together"; $\tau_b = -0.057 \pm 0.014$).[8] Anesthesiologists' mean supervision scores provided by residents and nurse anesthetists were not correlated with anesthesiologists' semi-annual clinical activity (multiple all $P > 0.65$).[4] A very active clinician can provide ineffective supervision, and a less active clinician can be very effective.[4] Supervision served as an independent contributor to the value that an individual anesthesiologist added to the care of the patient[4]
5. The most common supervision score provided by nurse anesthetists was 4.0 ($P < 0.0001$), indicating that all of the questions were considered important, including those related to teaching.[9] Anesthesiologist supervision scores provided by residents are even greater ($P < 0.0001$).[8,9] The pairwise differences by anesthesiologist are also significantly greater than zero ($P < 0.0001$)[8,9]
6. All residents evaluated all anesthesiologists' supervision during a study performed during a single weekend, such that each resident was in one class (e.g., "CA-1", "CA-2", etc.).[7] There was no association between residents' perception of supervision by anesthesiologists that met expectations and years since the start of training ($P = 0.77$).[5] There were very small differences among classes (mean differences ≤ 0.07 units).[7] Thus, "residents" can be treated as a single group, regardless of total years of clinical experience.

Table 2 continued

7. Mean resident scores for anesthesiologist's supervision were correlated with mean resident choice of the anesthesiologist to care for their family (Kendall's $\tau_b = +0.77$; $P < 0.0001$),[7] mean resident evaluations of the anesthesiologist's clinical teaching ($\tau_b = +0.87$; $P < 0.0001$),[7] and mean nurse anesthetist scores for anesthesiologist's supervision ($\tau_b = +0.36$; $P < 0.0001$ among all anesthesiologists and $\tau_b = +0.51$; $P < 0.0001$ among those with 15 raters of each type).
8. When the supervision instrument was applied to departments[12] (Tables 2.14 and 2.15), the internal consistency (Cronbach's alpha) of the scale was 0.909 ± 0.007. Convergent validity was based on a positive correlation between supervision and variables related to safety culture (all $P < 0.0001$): "Overall perceptions of patient safety", "Teamwork within units", "Non-punitive response to errors", "Handoffs and transitions", "Feedback and communication about error", "Communication openness", and the rotation's "overall grade on patient safety".[12] Convergent validity was based on significant negative correlation with variables related to the rater's burnout (all $P < 0.0001$): "I feel burnout from my work", "I have become more callous toward people since I took this job", and "errors with potential negative consequences to patients (that you have) made and/or witnessed".[12] Among these variables, supervision was most closely predicted by the same one variable using multiple types of regression trees: "Teamwork within (the rotation)" (e.g., "When one area in this rotation gets busy, others help out").[12] Discriminant validity was based on absence of rank correlation of supervision score with characteristics of raters and programs (all $P > 0.10$): age, hours worked per week, sex, promptness of survey response, number of survey raters from the department, and rotation (specialty) (as random effect)[12]
9. There was no significant association between anesthesiologist supervision score and the number of occasions that a resident rater had worked with the anesthesiologist, based on billing data (by patients, $\tau_b = +0.01$; $P = 0.71$ and by days, $\tau_b = -0.01$; $P = 0.46$)[7]
10. Among anesthesia residents, "the mean ± standard deviation of staff supervision scores that meets expectations", neither "exceeds expectations" nor is "below expectations" was 3.40 ± 0.30.[5] "Most … residents (94%) perceived that supervision that met their expectations was at least frequent (i.e., a score ≥ 3.0)" ($P < 0.0001$).[5] These values were greater than for nurse anesthetists ($P < 0.0001$)[5]
11. Anesthesia departments can measure individual anesthesiologists' supervision with high reliability (i.e., mean score is known with precision) when supervision scores are provided by at least nine different resident raters per anesthesiologist.[7] Monitoring is done by taking each of the raters' mean supervision scores for the anesthesiologist and weighting them equally (i.e., treating each rater's mean as a single observation)[8,15]
12. With residents' evaluations of anesthesiologists, mean supervision scores differed among anesthesiologists based on generalizability analysis ($P < 0.0001$)[7]
13. Anesthesiologist performance can be monitored daily using Bernoulli cumulative sum (CUSUM) control charts.[15] A reasonable threshold for low scores is < 3.0 for residents.[15] The true positive detection of anesthesiologists with incidences of low scores greater than the chosen "out-of-control" rate was 14/14.[15] The false-positive detection rate was 0/29.[15] Bernoulli CUSUM detection of low scores was within 50 ± 14 (median ± quartile deviation) days[15]

Table 2 continued

14. Anesthesia residents' mean scores for anesthesiologists' supervision for entire departments were significantly lower ($P < 0.0001$) than the mean scores for individual anesthesiologist's supervision.[13] The median ratio was 86% ± 1%. The correlation between departmental and mean (individual) anesthesiologists' scores was $\tau_b = 0.35 \pm 0.11$ ($P = 0.0032$).[13] When considering national survey results, individual anesthesiologists' supervisory performance needs to be greater.[13]
15. Anesthesia residents reporting mean supervision scores for their entire department (i.e., the mean of all anesthesiologists) that were < 3.00 (i.e., less than "frequent") reported anesthesiologists making more "mistakes that had negative consequences for the patient", with an accuracy (area under the curve) of 89% (99% confidence interval [CI], 77 to 95).[11] Supervision less than "frequent" (i.e., < 3.00) predicted "medication errors (dose or incorrect drug) in the last year" with an accuracy of 93% (99% CI, 77 to 98).[11] Among residents reporting overall supervision during the current rotation that was less than frequent (i.e., < 3.0) *vs* frequent, the 10th, 25th, 50th, 75th, 90th, and 95th percentiles of errors were 1 *vs* 1, 1 *vs* 1, 2 *vs* 2, 3 *vs* 2, 4 *vs* 3, and 6 *vs* 4, respectively ($P < 0.0001$).[12] There was no detected effect of resident burnout on numbers of reported errors while controlling for supervision (all $P > 0.138$ by different types of analyses)[12]
16. Nurse anesthetists' comments with (not) "see" or (not) "saw" and the theme "I did not see the anesthesiologist during the case(s) together" increased the odds of a nurse anesthetist providing a supervision score < 3 (odds ratio 48.2; $P < 0.0001$).[9] Many more such comments were made by nurse anesthetists than by residents ($P < 0.0001$).[9] Nevertheless, resident comments of insufficient anesthesiologist presence were associated with evaluation scores that were less than other evaluations with comments ($P < 0.0001$).[10] Each anesthesiologist with at least one such resident comment had lower mean scores than the other anesthesiologists ($P = 0.0071$)[10]
17. For both residents and nurse anesthetists, monitoring anesthesiologists' supervision and providing feedback resulted in greater scores by individual anesthesiologist.[4] For example, pairwise by anesthesiologist, the mean supervision scores provided by residents increased by 0.08 ± 0.01 points when equally weighting each anesthesiologist ($P < 0.0001$) and by 0.04 ± 0.02 points weighting by the precision of the difference ($P = 0.0011$).[4] Similarly, pairwise by anesthesiologist, the supervision scores provided by nurse anesthetists increased by 0.28 ± 0.02 points when equally weighting each anesthesiologist ($P < 0.0001$) and by 0.27 ± 0.02 points weighted by the precision of the difference ($P < 0.0001$).[4] Among nurse anesthetists, this was due principally to questions associated with teaching (e.g., "stimulate my clinical reasoning, critical thinking, and theoretical learning")[4]
18. Among anesthesia residents, evaluations of anesthesiologists with comments related to poor teaching had lower scores than the other evaluations with comments ($P < 0.0001$).[10] The anesthesiologists who each received a comment related to poor teaching had lower mean scores than the other anesthesiologists ($P < 0.0001$).[10] Each increase in the anesthesiologist's number of comments of poor-quality teaching was associated with a lower mean score ($P = 0.0002$).[10] Likewise, each increase in the anesthesiologist's number of resident comments of being disrespectful was associated with a lower mean supervision score ($P = 0.0002$).[10] A low supervision score (< 3.00; i.e., less than "frequent") had an odds ratio of 85 for disrespectful faculty behaviour ($P < 0.0001$)[10]

We have published previous summary tables in other papers on anesthesiologist supervision.[9,10]

References

1. *Dexter F*, *Ledolter J*, *Hindman BJ*. Quantifying the diversity and similarity of surgical procedures among hospitals and anesthesia providers. Anesth Analg 2016; 122: 251-63.
2. *Dexter F*, *Epstein RH*, *Dutton RP*, *et al*. Diversity and similarity of anesthesia procedures in the United States during and among regular work hours, evenings, and weekends. Anesth Analg 2016; 123: 1567-73.
3. *Epstein RH*, *Dexter F*. Influence of supervision ratios by anesthesiologists on first-case starts and critical portions of anesthetics. Anesthesiology 2012; 116: 683-91.
4. *Dexter F*, *Hindman BJ*. Quality of supervision as an independent contributor to an anesthesiologist's individual clinical value. Anesth Analg 2015; 121: 507-13.
5. *Dexter F*, *Logvinov II*, *Brull SJ*. Anesthesiology residents' and nurse anesthetists' perceptions of effective clinical faculty supervision by anesthesiologists. Anesth Analg 2013; 116: 1352-5.
6. *de Oliveira Filho GR*, *Dal Mago AJ*, *Garcia JH*, *Goldschmidt R*. An instrument designed for faculty supervision evaluation by anesthesia residents and its psychometric properties. Anesth Analg 2008; 107: 1316-22.
7. *Hindman BJ*, *Dexter F*, *Kreiter CD*, *Wachtel RE*. Determinants, associations, and psychometric properties of resident assessments of faculty operating room supervision. Anesth Analg 2013; 116: 1342-51.
8. *Dexter F*, *Ledolter J*, *Smith TC*, *Griffiths D*, *Hindman BJ*. Influence of provider type (nurse anesthetist or resident physician), staff assignments, and other covariates on daily evaluations of anesthesiologists' quality of supervision. Anesth Analg 2014; 119: 670-8.
9. *Dexter F*, *Masursky D*, *Hindman BJ*. Reliability and validity of the anesthesiologist supervision instrument when certified registered nurse anesthetists provide scores. Anesth Analg 2015; 120: 214-9.
10. *Dexter F*, *Szeluga D*, *Masursky D*, *Hindman BJ*. Written comments made by anesthesia residents when providing below average scores for the supervision provided by the faculty anesthesiologist. Anesth Analg 2016; 122: 2000-6.
11. *De Oliveira GS*, *Jr Rahmani R*, *Fitzgerald PC*, *Chang R*, *McCarthy RJ*. The association between frequency of self-reported medical errors and anesthesia trainee supervision: a survey of United States anesthesiology residents-in-training. Anesth Analg 2013; 116: 892-7.
12. *De Oliveira GS*, *Jr Dexter F*, *Bialek JM*, *McCarthy RJ*. Reliability and validity of assessing subspecialty level of faculty anesthesiologists' supervision of anesthesiology residents. Anesth Analg 2015; 120: 209-13.
13. *Hindman BJ*, *Dexter F*, *Smith TC*. Anesthesia residents' global (departmental) evaluation of faculty anesthesiologists' supervision can be less than their average evaluations of individual anesthesiologists. Anesth Analg 2015; 120: 204-8.
14. *Dexter F*, *Szeluga D*, *Hindman BJ*. Content analysis of resident evaluations of faculty anesthesiologists: supervision encompasses some attributes of the professionalism core competency. Can J Anesth 2017. DOI:10.1007/s12630-017-0839-7.
15. *Dexter F*, *Ledolter J*, *Hindman BJ*. Bernoulli cumulative sum (CUSUM) control charts for monitoring of anesthesiologists' performance in supervising anesthesia residents and nurse anesthetists. Anesth Analg 2014; 119: 679-85.
16. *Epstein RH*, *Dexter F*, *Patel N*. Influencing anesthesia provider behavior using anesthesia information management system data for near real-time alerts and post hoc reports. Anesth Analg 2015; 121: 678-92.

17. *O'Neill L*, *Dexter F*, *Zhang N*. The risks to patient privacy from publishing data from clinical anesthesia studies. Anesth Analg 2016; 122: 2017-27.
18. *Feldt LS*, *Woodruff DJ*, *Salih FA*. Statistical inference for coefficient alpha. Appl Psychol Meas 1987; 11: 93-103.
19. *Yamamoto S*, *Tanaka P*, *Madsen MV*, *Macario A*. Analysis of resident case logs in an anesthesiology residency program. A A Case Rep 2016; 6: 257-62.
20. *Doane DP, Seward LE*. Measuring skewness: a forgotten statistic? J Stat Educ 2011; 19 (2).
21. *Box GE*, *Cox DR*. An analysis of transformations. J R Stat Soc Series B Stat Methodol 1964; 26: 211-52.
22. *Dexter F*, *Wachtel RE*, *Todd MM*, *Hindman BJ*. The "fourth mission:" the time commitment of anesthesiology faculty for management is comparable to their time commitments to education, research, and indirect patient care. A A Case Rep 2015; 5: 206-11.
23. *Austin PC*, *Alter DA*, *Tu JV*. The use of fixed-and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. Med Decis Making 2003; 23: 526-39.
24. *Racz MJ*, *Sedransk J*. Bayesian and frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes. J Am Stat Assoc 2010; 105: 48-58.
25. *Yang X*, *Peng B*, *Chen R*, *et al*. Statistical profiling methods with hierarchical logistic regression for healthcare providers with binary outcomes. J Appl Stat 2014; 41: 46-59.
26. *Glance LG*, *Li Y*, *Dick AW*. Quality of quality measurement: impact of risk adjustment, hospital volume, and hospital performance. Anesthesiology 2016; 125: 1092-102.
27. *Dexter F*, *Hindman BJ*. Do not use hierarchical logistic regression models with low-incidence outcome data to compare anesthesiologists in your department. Anesthesiology 2016; 125: 1083-4.
28. *Baker K*. Determining resident clinical performance: getting beyond the noise. Anesthesiology 2011; 115: 862-78.
29. *Baker K*, *Sun H*, *Harman A*, *Poon KT*, *Rathmell JP*. Clinical performance scores are independently associated with the American Board of Anesthesiology Certification Examination scores. Anesth Analg 2016; 122: 1992-9.
30. *Whyte S*, *Cartmill C*, *Gardezi F*, *et al*. Uptake of a team briefing in the operating theatre: a Burkean dramatistic analysis. Soc Sci Med 2009; 69: 1757-66.
31. *Einav Y*, *Gopher D*, *Kara I*, *et al*. Preoperative briefing in the operating room: shared cognition, teamwork, and patient safety. Chest 2010; 137: 443-9.