Canadian Medical Education Journal

Major Contribution

Examining the accuracy of residents' self-assessments and faculty assessment behaviours in anesthesiology

Examen de l'exactitude des autoévaluations des résidents et des comportements d'évaluation des professeurs en anesthésiologie

Melinda Fleming,[1] Danika Vautour,[1] Michael McMullen,[1] Nicholas Cofie,[2] Nancy Dalgarno,[2] Rachel Phelan,[1] Glenio B Mizubuti[1]

[1]Department of Anesthesiology and Perioperative Medicine, Kingston Health Sciences Centre; [2]Faculty of Health Sciences, Queens University, Ontario, Canada.

Correspondence to: Dr. Glenio B. Mizubuti; email: Glenio.Mizubuti@Kingstonhsc.ca

Published ahead of issue: April 12, 2021; CMEJ 2021 Available at http://www.cmej.ca

https://doi.org/10.36834/cmej.70697.

Abstract

Background: Residents' accurate self-assessment and clinical judgment are essential for optimizing their clinical skills development. Evidence from the medical literature suggests that residents generally do poorly at self-assessing their performance, often due to factors relating to learners' personal backgrounds, cultures, the specific contexts of the learning environment and rater bias or inaccuracies. We evaluated the accuracy of anesthesiology residents' self-assessed Global Entrustment scores and determined whether differences between faculty and resident scores varied by resident seniority, faculty leniency, and/or year of assessment.

Methods: We employed variance components modeling techniques and analyzed 329 pairs of faculty and self-assessed entrustment scores among 43 faculty assessors and 15 residents. Using faculty scores as the gold standard, we compared faculty scores with residents' scores ($x_{i(faculty)}-x_{i(resident)}$), and determined residents' accuracy, including over- and under-confidence.

Results: The results indicate that residents were respectively over- and under-confident in 10.9% and 54.4% of the assessments but more consistent in their individual self-assessments (*rho* = 0.70) than faculty assessors. Faculty scores were significantly higher (α = 0.396; z = 4.39; $p < 0.001$) than residents' self-assessed scores. Being a lenient/dovish (β = 0.121, z = 3.16, $p < 0.01$) and a neutral (β = 0.137, z = 3.57, $p < 0.001$) faculty assessor predicted a higher likelihood of resident under-confidence. Senior residents were significantly less likely to be under-confident compared to junior residents (β = -0.182, z =-2.45, $p < 0.05$). The accuracy of self-assessments did not significantly vary during the two years of the study period.

Conclusions: The majority of residents' self-assessments were inaccurate. Our findings may help identify the sources of such inaccuracies.

Abstract

Contexte : L'autoévaluation adéquate et le jugement clinique des résidents sont essentiels pour optimiser le développement de leurs compétences cliniques. Les données probantes de la littérature médicale indiquent que les résidents ont généralement du mal à s'auto-évaluer, souvent en raison de facteurs liés à leur passé personnel, à la culture, aux contextes spécifiques de l'environnement d'apprentissage et aux préjugés ou inexactitudes des évaluateurs. Nous avons évalué l'exactitude des scores d'autoévaluation par échelles de confiance globalepar des résidents en anesthésiologie et déterminé si les différences entre les scores des enseignants et des résidents variaient en fonction du niveau de formation des résidents, de l'indulgence des enseignants ou de l'année d'évaluation.

Méthodes : Nous avons utilisé des techniques de modélisation des composantes de la variance et analysé 329 paires de scores de confiance des enseignants et d'autoévaluation avec la participation de 43 évaluateurs du corps professoral et 15 résidents. Prenant les scores des enseignants comme référence, nous avons comparé les leurs avec ceux des résidents ($x_{i(enseignant)}-x_{i(résident)}$), et déterminé l'exactitude chez les résidents, y compris l'excès et le manque de confiance.

Résultats : Les résultats indiquent que les résidents étaient trop confiants dans 10,9 % des évaluations, et pas assez confiants dans 54,4 % des cas, mais qu'ils étaient plus cohérents dans leurs autoévaluations (rho = 0,70) que ne l'étaient les enseignants. Les scores du corps professoral étaient significativement plus élevés (α = 0,396 ; z = 4,39 ; $p < 0,001$) que les scores d'autoévaluation des résidents. Le fait d'être un évaluateur indulgent (β = 0,121, z = 3,16, $p < 0,01$) et neutre (β = 0,137, z = 3,57, $p < 0,001$) prédisait une probabilité plus élevée de sous-confiance des résidents. Les résidents seniors étaient significativement moins susceptibles de manquer de confiance en eux que les résidents juniors (β = -0,182, z = -2,45, $p < 0,05$). L'exactitude des autoévaluations n'a pas varié de manière significative au cours des deux années de la période d'étude.

Conclusions : La majorité des autoévaluations des résidents étaient inexactes. Nos résultats peuvent aider à identifier les sources de ces inexactitudes.

Introduction

The accuracy of residents' self-assessment and clinical judgment is essential for resident development, clinical improvement, and patient safety. Self-assessment has been described as an ingrained habit or trait of reflective individuals.[1] It reflects one's ability to make judgments regarding their own competence, and is critical for learners' self-regulation, and professional development.[2-5] As a skill, self-assessment requires deliberate training and practice.[6] Boud noted that an effective self-assessment requires residents to have a clear understanding of the performance standards and the criteria for measuring them, and be able to use these criteria to make accurate judgments about their own performance.[7]

Evidence from the medical literature suggests that residents generally do poorly at self-assessing their performance as compared to faculty evaluation.[5,8-10] Also, research suggests that medical learners who appear over-confident tend to overestimate their skills and rate themselves highly while under-confident learners tend to underestimate their ability to accurately self-assess, thereby performing poorly in self-assessments.[8,11-14] Other factors often cited for the inaccuracies in self-assessment include: personal background and culture; the contexts in which learners work; and the specific learning environment.[10,15-17] Accuracy of self-assessment has been shown to remain stable or improve (within a range of subjects) over time, although it may decline with unfamiliar tasks.[18,19] Also, self-assessment in higher education is generally known to be more accurate as seniority increases due to senior trainees' increased confidence and experience in learning and understanding the rudiments of self-assessment as a skill.[2,4,6,7,14,20,21]

The accuracy of learners' assessments is affected by rater bias and/or inaccuracies which may be systematic or non-systematic.[22] A systematic bias reveals a *rater effect* where there is a systematic variance in performance ratings associated with the rater and not with the actual performance of the ratee.[23] On the other hand, a non-systematic bias refers to a *rater error* where an individual's idiosyncrasies are associated with the random interaction effects of the rater, the testing situation, and the ratee.[24] Assessor bias is an important source of rater error in faculty assessment of residents, and reduces the reliability and validity of such assessments.[25-27] Rater errors include those arising from rater variance, extreme rater stringency, and leniency biases.[27] Assessor leniency bias may emanate from the error in rating trainees' abilities significantly higher than their performance merits, or an assessor's failure to fail.[28] McManus, Thompson and Mollon,[29] for instance, described low inter-rater reliability of assessment scores where multiple assessors differed in their opinion on a performance, such as in an Objective Structured Clinical Examination (OSCE). Others, such as Pell et al.[30] and Bartmann et al.[22] have identified extreme "hawk" and "dove" behaviours in a post-hoc analysis of OSCE scores. Indeed, the 'hawk/dove' problem has been known qualitatively since the early 1900s. For example, Osler identified examiners who were extremely reluctant to pass any students, examiners who were dismayed by the thought of failing any students, and a "large group of sensible examiners" who performed their task in a reasonable and fair manner.[31]

The current study is important in that to the best of our knowledge it is the first attempt to uniquely use longitudinal panel data to directly examine the impact of faculty leniency/stringency on the accuracy of residents' self-assessments outside the context of a summative assessment (e.g., OSCE). Unlike previous studies that tend to attribute poorly performed self-assessments mainly (or solely) to the learners, we argue (and demonstrate empirically) that some of the observed inaccuracies in residents' self-assessments may be explained by the extent of faculty assessor leniency or stringency. We therefore hypothesized that faculty assessor leniency would be associated with higher resident accuracy. We also argue that senior residents, being more experienced and confident,[2,4,14,20,21] would likely be more accurate in their self-assessments than junior residents. Further, while previous research examining residents' accuracy has been primarily based on snapshots of self-assessment activities,[18,19] the results of investigations (like the present one) where data were collected longitudinally will arguably be more robust and reliable. We therefore evaluated the accuracy of residents' self-assessed global entrustment scores in comparison to faculty assessed scores, and whether the differences observed varied significantly by residents' seniority, faculty leniency, and/or the year of assessment.

Methods

Data collection

Data for this study were derived from an analytic sample of 329 pairs of global entrustment assessment scores provided by 15 residents and 43 faculty assessors in the

anesthesiology residency training program at our medical institution. All assessments in the analytic sample were completed in 2017 and 2018. The study protocol including information on participants' consent was reviewed and approved by the Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (TRAQ# 6020176). Assessment ratings were based on a 5-point anchored Likert scale with higher values indicating better overall entrustment ability of residents. Specific domains and definitions for a given score are shown in Table 1. Each score on the scale is anchored upon a specific domain/descriptor.[32] For example, a faculty assessor providing a global entrustment assessment score of 5 agrees to the statement that "I did not need to be there: Completely independent, understands risks and performs safely, insightful, pre-emptive and proactive, ready for practice." All missing assessment scores were assumed to be missing at random.

Table 1. Definitions of faculty and residents' assessment scores

Rating	Definition
1	"I had to do it completely": Requires complete hands on guidance, did not do, or not given the opportunity to do.
2	"I had to intervene or talk the resident through": Able to perform task but needs constant direction.
3	"I had to prompt them from time to time": Demonstrates some independence but requires intermittent direction.
4	"I needed to be in the room just in case": Independent but unaware of all risks and still requires supervision or consultation for safe practice.
5	"I did not need to be there": Completely independent, understands risks and performs safely, insightful, pre-emptive and proactive, ready for practice.

Outcome variable

The main outcome variable of interest is accuracy of self-assessments. Following Gordon's[1] approach, we defined and measured accurate assessments as the degree to which residents' self-assessed scores aligned with faculty assessment scores. Despite some criticism, the use of subjective faculty assessments of residents as the gold standard has been described as a good proxy for measuring clinical competence.[33] Thus, using faculty scores as the gold standard, we compared faculty scores with residents' scores ($x_{i(faculty)} - x_{i(resident)}$) and determined whether residents were accurate, over-confident, or under-confident (see Table 2). When residents' scores were equal to faculty scores, they were classified as accurate. When residents and faculty scores differed, such scores were described as over-confident ($x_{i(faculty)} - x_{i(resident)} < 0$) or under-confident ($x_{i(faculty)} - x_{i(resident)} > 0$) respectively. Thus, on a continuum, higher values on the accuracy scale represent under-confidence whereas lower values reflect over-confidence.

Independent variables

We used faculty assessor leniency (doves) or stringency (hawks) as the main independent variable. We computed assessor "hawkishness" or "dovishness" by comparing a faculty assessor's average rating for a resident to the average rating assigned by all other faculty assessors for that resident. Using the Gaussian distribution and 95% confidence level, we transformed the resulting differential scores into z scores, and defined assessors whose z scores were below -1.96 as hawks, and those with scores above +1.96 as doves (see Table 3). We also controlled for the seniority of residents, period of assessment, and year in which the assessments were completed. Residents in their fourth and fifth years of training were classified as seniors, while those with up to three years of training were classified as juniors.

Analytical strategy

We analyzed data for all variables of interest using descriptive and multilevel inferential statistical techniques. Given the evidence of clustering and the panel nature of the data with residents self-assessing and being assessed multiple times per year, we employed variance components modeling techniques, and estimated random intercept linear regression models which allow for modeling the structure of the within-panel correlation.[34-37] This procedure allows the regression parameter estimates to be unbiased and efficient. Stata statistical software was used to perform the analysis.[38]

Results

Descriptive statistics examining the distribution of resident and faculty assessment scores

Faculty assessors assigned relatively higher global entrustment scores (≥ 4) than residents whose assessment scores were mostly clustered around scores of 2.5, 3.5, and 4 (see Figure 1). Residents' self-assessment scores were moderately correlated with faculty assessment scores (r = 0.51, p < 0.001). Table 2 demonstrates that residents were over- and under-confident in 10.9% and 54.4% of the assessments, respectively, with approximately a third of the assessments being accurate (34.7%).

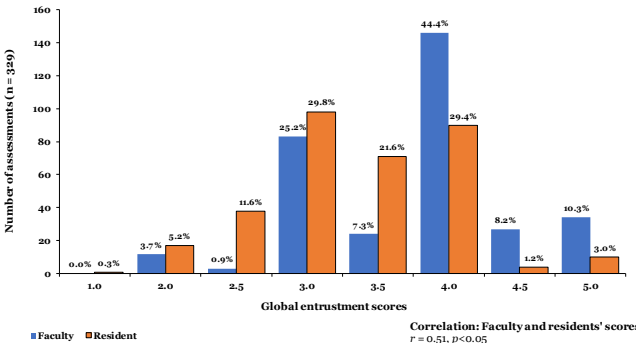

Figure 1. Distribution of faculty and residents' assessment scores

Table 2. Descriptive statistics for panel data examining differences in faculty and residents' assessments

	Overall sample		Between residents		Within residents
Δ(Faculty-resident scores)	*N*	%	*N*	%	%
-2.0	1	0.30	1	6.67	3.57
-1.5	2	0.61	2	13.33	2.05
-1.0	6	1.82	3	20.00	6.15
-0.5	27	8.21	10	66.67	15.37
0.0	114	34.65	15	100.00	40.91
+0.5	78	23.71	14	93.33	21.1
+1.0	71	21.58	13	86.67	23.01
+1.5	16	4.86	7	46.67	8.57
+2.0	11	3.34	6	40.00	6.51
+2.5	2	0.61	1	6.67	8.7
+3.0	1	0.30	1	6.67	4.35
Total (*N* = 15)	329	100	73	486.67	20.55

Notes: 1] (a) Accurate scores: 34.65%; (b) Under-confident scores: 54.41%; (c) Over-confident scores: 10.94%
2] Positive (+) values represent under-confidence; negative (-) values represent over-confidence; and 0 (zero) represents accurate assessments.
3] Between residents' scores interpretation (e.g.,): Among the 15 residents, all of them (100%) were accurate in at least one self-assessment.
4] Within residents' scores interpretation (e.g.,): Of these 15 residents (100%) who were accurate in at least one self-assessment, they were on average only about 41% accurate in all their assessments.

Table 2 further shows that all of the 15 residents (100%) were accurate in at least one of the assessments they completed, yielding an average of 40.9% accuracy rate in these particular self-assessments. Additionally, 46.7% and 13.3% of the residents scored 1.5 less and 1.5 more than faculty assessors, respectively, in at least one of their assessments. Most residents (86.7%) scored 1.0 in excess of faculty scores, averaging 23.0% of all of their assessments. Table 4 shows that the majority of accurate (72.2%), over- (97.4%), and under-confident (92.7%) scores were assessed by neutral (neither stringent nor lenient) faculty (*n* = 40). Stringent (or hawkish) faculty (*n* = 2) assessed 27.8% of over-confident scores as well as 3.9% of under-confident sores. None of the over-confident or accurate scores were assessed by lenient (or dovish) assessors. Dovish faculty (*n* = 1) assessed 3.4% of all under-confident scores. According to Table 3, the majority of assessments (92.0%) were completed by neutral faculty and variability in assessment scores varied significantly among faculty assessors (*Wo* = 14.2, df(2, 236), $p < 0.001$). The largest variability occurred among stringent faculty (SD = 1.15), whereas lenient faculty assessors produced the least variance, with scores ranging from 2 to 5, and 4.5 to 5, respectively. Stringent faculty assigned a greater number of scores below 3 (*n* = 9) as compared with neutral (*n* = 6) and lenient faculty (*n* = 0).

Table 3. Distribution of faculty global assessment scores by type of assessor

	N	# of assessments	Mean /SD	Range of scores	# of Scores < 3
Faculty assessor type					
Stringent	2	20 (6.08%)	3.05 (1.15)	2 - 5	9
Neutral	40	303 (92.09%)	3.8 (0.65)	2 - 5	6
Lenient	1	6 (1.80%)	4.67 (0.26)	4.5 - 5	0
N	43	329			15

Note: Levene's robust test statistic (W0) for the equality of variances between the assessor groups = 14.203, df(2, 236), Prob > F =0.00000.

Table 4. Distribution of over-confident, under-confident and accurate scores by type of assessor

	Over-confidence scores	Accurate scores	Under-confidence scores
Faculty assessor type	%	%	%
Stringent	27.78	2.63	3.91
Neutral	72.22	97.37	92.74
Lenient	0.00	0.00	3.35
N	36	114	179

Testing the alignment between resident and faculty assessment scores

We assessed whether residents' self-assessed global entrustment scores differed significantly from faculty-assessed scores. Results from a random intercept model testing the null hypothesis of no difference between faculty and residents' scores shows that faculty

assessment scores were on average, 0.4 points higher (α = 0.396, z = 4.39, p < 0.001) than residents' self-assessed scores (see Table 5). Intra-class correlations presented in Table 6 suggest that individual residents tended to be more consistent/reliable (*rho* = 0.70) in their assessments than faculty assessors (*rho* = 0.20). Accordingly, the amount of variance observed within residents' self-assessments (θ^2 = 0.40, CI = 0.376 – 0.433) was much lower than the within faculty variance observed (θ^2 = 0.64, CI = 0.598 – 0.692). On the other hand, the between faculty assessment scores appear to be more similar with less variability (ψ^2 = 0.32, CI = 0.236 – 0.444) compared to the between resident self-assessment scores (ψ^2 = 0.61, CI = 0.423 – 0.882).

Table 5. Random intercept model examining differences in faculty and residents' scores

Δ(Faculty-resident scores)	Coef.	Std. err	Z	p >|z|	95% Confidence interval	
Constant (α)	0.396		4.39	0.00	0.219	0.572
Random intercepts						
Between residents variance $(\psi)^2$	0.306	0.070			0.196	0.479
Within residents variance $(\phi)^2$	0.618	0.025			0.571	0.668
Rho (intra-class correlation)	0.197	0.074			0.085	0.371

Note: This is a null or univariate random intercept model (without any independent variables) testing whether differences in faculty and residents' scores are equal to zero.

Table 6. Random intercept model examining faculty and residents' assessment scores

Residents' self-assessment scores		Residents' Model		
	Coef.	Std. err	z	p >|z|
Constant (α)	3.58	0.16	22.39	0.000
Random intercepts				
Between residents variance $(\psi)^2$	0.61	0.14		
Within residents variance $(\phi)^2$	0.40	0.01		
Rho (intra-class correlation)	0.70	0.08		
Faculty assessment scores		Faculty Model		
Constant (α)	3.82	0.06	61.02	0.000
Between residents variance $(\psi)^2$	0.32	0.05		
Within residents variance $(\phi)^2$	0.64	0.02		
Rho (intra-class correlation)	0.20	0.05		

The effects of faculty leniency, resident seniority, and year of assessment on accuracy of residents' self-assessments

We also assessed whether differences in faculty and residents' scores vary significantly by faculty leniency, seniority of residents, and/or year of assessment. Table 7 shows that compared with hawkish faculty assessors, residents were significantly more likely to be under-confident if they were assessed by dovish (β = 0.121, z = 3.16, p < 0.01) or neutral faculty assessors (β = 0.137, z = 3.57, p < 0.001). Senior residents were significantly less likely to be under-confident compared with junior residents (β = -0.182, z = -2.45, p < 0.05). The year of assessment had no effect on residents' accuracy (β = 0.047, z = 0.58, p > 0.05).

Discussion

Overview of key findings and implications

This study examined the accuracy of residents' self-assessed global entrustment scores and determined whether self-assessed scores significantly differed from faculty assessed scores. We also examined whether differences in faculty and residents' scores varied by residents' seniority, faculty stringency/leniency, and/or year of assessment. Using faculty scores as the gold standard,[33] our findings highlight that two out of three resident self-assessments were inaccurate. Also, self-assessed scores were only moderately correlated with faculty scores, indicating a limited agreement between resident and faculty assessments.

Faculty scores were found to be significantly higher than residents' self-assessed scores, indicating that residents are often under-confident in their assessments. We found this observation to be robust even after controlling for relevant covariates in the multivariate analysis. Indeed, evidence of under-confidence was present in more than half of the assessments, and it is instructive to determine whether residents were truly under-confident or whether they were under-confident merely because they were assessed by a more lenient (or dovish) faculty assessor. If the latter was true, we would expect that the majority of the assessments performed by lenient assessors would be classified as under-confident scores. Conversely, we would expect that a greater proportion of assessments performed by stringent assessors would produce over-confident scores. Finally, we would expect faculty leniency to predict a higher likelihood of resident under-confidence in a multivariable context.

Robustness and dynamics of key findings

The evidence from our study demonstrates that when self-assessments were compared with scores provided by a lenient faculty assessor, all were classified as under-confident (see Table 8). This suggests probably that

residents may have been classified as under-confident potentially because they may have been assessed by a more lenient assessor. Similarly, assessor leniency was found to be independently and positively associated with resident under-confidence in the multivariate analysis (see Table 7). However, only half of the assessments provided by stringent faculty assessors were deemed over-confident scores (see Table 8). Alternatively, it may be that under-confident residents were simply residents that did not want to appear over-confident because they were actually aware of the nature of the ongoing study. Residents may also have been aware of the faculty rater's hawkish/dovish bias, which may have influenced their self-assessment scores.

Table 7. Multivariate random intercept model examining differences in faculty and residents' scores

	Standardized		
Δ(Faculty-resident scores)	Coef.	Std. err	z
Constant (α)	0.45***	0.08	5.48
Faculty assessor type			
Stringent (Ref. category)			
Neutral	0.14***	0.04	3.57
Lenient	0.12**	0.04	3.16
Resident Seniority			
Junior (Ref. category)			
Senior	-0.18*	0.07	-2.45
Year of assessment	0.05	0.08	0.58
Period of assessment			
First quarter (Ref. category)			
Second quarter	0.01	0.15	0.13
Third quarter	0.08	0.14	0.58
Random intercepts			
Between residents variance $(\psi)^2$	0.27	0.06	
Within residents variance $(\varnothing)^2$	0.60	0.02	
Rho (intra-class correlation)	0.16	0.06	

Note: All predictor variables were standardized.
Statistical significance: *p<0.05, **p<0.01, ***p<0.001

Table 8. Percentage distribution of the number of assessments performed by faculty assessors

	Faculty assessor type		
	Stringent	Neutral	Lenient
Accuracy of assessments	%	%	%
Over-confidence scores	50.00	8.58	0.00
Accurate scores	15.00	36.63	0.00
Under-confidence scores	35.00	54.79	100.00
N	20	303	6

Interestingly, stringent assessors appeared to be lenient in more than a third of their assessments, indicating some fluidity in faculty hawkishness within the current sample. We again speculate that a Hawthorne effect may have played a role, as all assessors were aware of our data collection. Our data revealed that as opposed to hawks, who assigned a wider range of scores (range: 2 – 5), doves had a tendency to use a narrow range of scores (range: 4.5 – 5) (see Table 3). Comparatively, residents produced the largest variability in the assessments (range: 1 – 5) (see Table 6). Additionally, it is worth noting that hawks were the only assessors who used round numbers as opposed to half scores (e.g., 3.5, 4.5, etc.) in their assessments. Variance in the use of a range of scores may be associated with better quality assessments. The use of a greater range of score by stringent faculty indicate a more descriptive assessment as hawks used 4 out of 5 markers on the 5-point Likert scale, while doves used only 2 out of 10 markers on the self-created 9-point Likert scale, with their use of half scores. Faculty assessors' use of half scores (though not part of the global entrustment scale design) may have been facilitated by the paper-based nature of the assessment records as opposed to computer-based assessments where only round number of scores may be permitted. Assigning half scores may reflect either a leniency or stringency bias toward residents whose performance, in reality, may not merit certain round number scores.

Also, the finding that faculty assessors were neutral in their assessments produced ratings that had an inclination toward resident under-confidence is an important one. Out of the 303 assessments done by neutral faculty, only 6 (1.98%) had an entrustment score of less than 3. These dynamics in faculty ratings on the global entrustment scale provide a clear opportunity for faculty development as there seems to be an overall tendency towards lenient scoring. Further research may be needed to understand why faculty neutrality may be associated with resident under-confidence.

Faculty development has been proposed as a way of improving rater bias,[39,40] although such interventions could still result in faculty scores that are more lenient[41] or more stringent.[42] It has been suggested that residency programs can identify a subset of faculty dedicated to learning and performance assessments, and make them responsible for the majority of resident assessments.[41,42] Yet, others point out that this strategy might be less effective since medical raters may be unresponsive to

professional development (e.g., Cook et al.).[43] Cook and colleagues' observation raises an important question regarding whether there is still value in implementing faculty development strategies that may not yield the desired results.[43]

From a resident development perspective, residents may be interested in knowing: "Who is actually assessing me and what inherent biases does the assessor have?" Also, members of a competency or promotion committee in a residency program should be interested in knowing: "Who assessed this trainee, and was the assessor unbiased or objective?" Answers to these difficult questions can have significant implications for the quality of assessments and overall residency training. For instance, identifying and correcting an extremely over-confident or a less confident resident may be difficult if there is a prevailing and systematic leniency bias or culture amongst faculty assessors.

Our findings further demonstrate that consistent with extant research, resident seniority is associated with less under-confidence.[2,4,21] The inaccuracies observed between resident self-assessments and faculty scores were consistently observed throughout the two years of our study. It may be that the two-year period within which the assessments were completed was too short to allow any meaningful differences to be detected as an individual resident increased in seniority.

Value of current findings

To our knowledge, the current study is the first longitudinal study directly examining the impact of faculty leniency/stringency on residents' accuracy in self-assessments outside of the context of an OSCE. Our findings provide medical educators the unique opportunity to better understand the accuracy of resident and faculty assessments of performance in a dynamic environment in which complex clinical care is provided. Unlike other residency training settings, the nature of anesthesiology typically limits performance assessments to a single context (surgical cases in the operating room) where faculty directly observe and interact with a single learner. This training model should provide the faculty assessor the opportunity to better judge the clinical competencies of the trainee. As Tavares and Eva[44] note, other clinical training settings are typically busy and distracting to the clinical environment in which supervisors are asked to assess learners, and this may cause cognitive overload and impact the accuracy of the assessments.[45]

Limitations

Limitations of our study include those related to data collection, data quality, the measurement of the accuracy of assessments, the unequal number of assessments completed by faculty, and the low number of lenient (n = 1) and stringent (n = 2) assessors. Both faculty and residents did not fully provide their respective entrustment scores for every global entrustment assessment completed and this accounted for about 32.4% reduction in the analytic sample size. Thus, we analyzed only data with complete resident information, assuming such data were missing at random. Given the possibility that such data may have been missing for reasons other than random omission, a cautionary generalization of the current findings is urged. The addition of qualitative interviews to better understand faculty and residents' rationale for not fully completing their respective self-assessments could have improved our study design. Also, as previously noted, residents were more likely to be biased in their self-assessments since they were aware that their assessments were going to be a part of a research project.

Though conceptually reasonable and analytically unambiguous, the use of the deltas between faculty and residents' assessment scores to define the accuracy of self-assessed scores may be crude and could mask real differences between faculty and resident scores. Ideally, more data may be needed to establish statistically significant thresholds at which a residual score may be considered a significant change. This way, significant differences in scores such as those greater ± 1 (e.g., delta >1) may be used to trigger an academic meeting to explore the reasons behind the difference. As Table 2 shows, about 90% of the assessments were either accurate or varied by a unit of ± 1, and with more data, it would be plausible to further investigate whether the use of statistically tested thresholds may produce better alternative outcomes of accuracy. Similarly, because some faculty assessed fewer residents than other faculty assessors, the use of deltas to define assessor leniency or stringency could be potentially affected by the unequal number of assessments that faculty assessors completed. It is equally important to note that the low number of lenient and stringent assessors discovered in the analysis could affect the robustness of the results and thus bias our ability to make causal inferences. Finally, alternative measures such as the use of group means of self-assessments across domains of interest have been

described in the literature and can be used within specific contexts and guidelines.[46-48]

Conclusion

Our study demonstrates that the majority of residents' self-assessments were inaccurate. Senior residents were less often under-confident than junior residents, and accuracy did not improve over time. Stringent assessors were not always hawkish in their assessments as evidenced in the wider range of their entrustment scores. They were actually lenient in more than a third of the assessments they performed, demonstrating fluidity and effectiveness in their assessments. We identified an overall tendency toward faculty leniency in resident assessments and there is evidence that faculty development could improve this rater-based assessment bias. Our findings and the ongoing use of self-assessments may provide educators the opportunity to identify profiles of under- and over-confident residents to inform programs aimed at improving competency-based entrustment assessments.

Conflicts of Interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this manuscript.

Funding: This study was funded by the Queen's University Faculty of Health Sciences Postgraduate Medical Education Special Purpose Grant.

References

1. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med*. 1991; 66(12):762-769. https://doi.org/10.1097/00001888-199112000-00012
2. Ganni S, Jakimowicz JJ. "Reflection-before-practice" improves self-assessment and end-performance in laparoscopic surgical skills training. *J Surg Educ*. 2017; 75(2): 527-533. https://doi.org/10.1016/j.jsurg.2017.07.030
3. Perera J, Perera J, Abdullah J, Lee N. Training simulated patients: evaluation of a training approach using self-assessment and peer/tutor feedback to improve performance. *BMC Med Educ*. 2009; 9(1): 37. https://doi.org/10.1186/1472-6920-9-37
4. Plant JL, van Schaik SM, Sliwka DC, Boscardin CK, O'Sullivan PS. Validation of a self-efficacy instrument and its relationship to performance of crisis resource management skills. *Adv Health Sci Educ*. 2011;16(5):579-590. https://doi.org/10.1007/s10459-011-9274-7
5. Violato C, Lockyer J. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: Implications for self-directed learning. *Adv Health Sci Educ*. 2006; 11(3):235-244. https://doi.org/10.1007/s10459-005-5639-0
6. Fuhrmann B, Weisburg M. Self-evaluation. In: Morgan M, Irby D, eds. *Evaluating clinical competence in the health professions*. St Louis Missouri: Mosby; 1978. pp. 139-150.
7. Boud D. *Enhancing Learning through Self-assessment*. London: Kogan Page Ltd;1995.
8. Chow I, Nguyen VT, Losee JE, et al. Milestones in plastic surgery: Attending assessment versus resident assessment. *Plast Reconstr Surg*. 2019; 143(2): 425e-432e. https://doi.org/10.1097/PRS.0000000000005214
9. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. *J Am Med Assoc*. 2006; 296 (9): 1094-102. https://doi.org/10.1001/jama.296.9.1094
10. Eva KW, Regehr G. Self-assessment in the health professions. *Acad Med*. 2005; 80 Supp: S46-54. https://doi.org/10.1097/00001888-200510001-00015
11. Davies JG, Ciantar J, Jubraj B, Bates IP. Use of a multisource feedback tool to develop pharmacists in a postgraduate training program. *Am J Pharm Educ*. 2013; 77(3):1-7. https://doi.org/10.5688/ajpe77352
12. Pandey VA, Wolfe JHN, Black SA, Cairols M, Liapis CD, Begqvist D. Self-assessment of technical skill in surgery: The need for expert feedback. *Ann R Coll Surg Engl*. 2008; 90(4):286-290. https://doi.org/10.1111/j.1478-4408.1974.tb03207.x
13. Tedesco MM, Pak JJ, Harris, EJ, Krummel TM, Dalman RL, Lee JT. Simulation-based endovascular skills assessment: The future of credentialing? *J Vas Surg*. 2008; 47(5):1008-1013. https://doi.org/10.1016/j.jvs.2008.01.007
14. Vnuk A, Owen H, Plummer J. Assessing proficiency in adult basic life support: student and expert assessment and the impact of video recording. *Med Teach*. 2006; 28(5):429-434.
15. Brydges R, Butler D. A reflective analysis of medical education research on self-regulation in learning and practice. *Med Educ*. 2012; 46(1): 71-79. https://doi.org/10.1111/j.1365-2923.2011.04100.x
16. Eva KW, Regehr G. Knowing when to look it up: A new conception of self-assessment ability. *Acad Med*. 2007; 82 Supp: S81-84. https://doi.org/10.1097/ACM.0b013e31813e6755
17. Sargeant J, Armson H, Chesluk B, et al. The processes and dimensions of informed self-assessment: a conceptual model. *Acad Med*. 2010; 85(7): 1212-1220. https://doi.org/10.1097/ACM.0b013e3181d85a4e
18. Boud D, Lawson R, Thompson DG. Does student engagement in self-assessment calibrate their judgement over time? *Ass Eval High Educ.* 2013; 38(8): 941-956. https://doi.org/10.1080/02602938.2013.769198

19. Fitzgerald JT, White CB, Gruppen LD. A longitudinal study of self-assessment accuracy. *Med Educ.* 2003; 37(7):645-649. https://doi.org/10.1046/j.1365-2923.2003.01567.x
20. Boud D, Falchikov N. Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *High Educ*.1989; 18(5): 529-549. https://doi.org/10.1007/BF00138746
21. Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. Self-assessment of performance among surgical trainees during simulated procedures in a simulated operating theater. *Am J Surg.* 2006; 192(1):11-8. https://doi.org/10.1258/135581906777641640
22. Bartman I, Smee S, Roy M. A method of identifying extreme OSCE examiners. *Clin Teach*. 2013; 10(1): 27-31. https://doi.org/10.18546/RET.31.1.06
23. Scullen SE, Mount MK, Goff M. Understanding the latent structure of job performance ratings. *J Appl Psychol*. 2000; 85(6): 956–970. https://doi.org/10.1037/0021-9010.85.6.956
24. Harasym PH, Woloschuck W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ*. 2008;13:617–632. https://doi.org/10.1007/s10459-007-9068-0
25. Govaerts MJB, Van Der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Adv Health Sci Educ*. 2007; 12(2):239-260. https://doi.org/10.1007/s10459-006-9043-1
26. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Acad Med*. 2011; 86: Supp S1-7. https://doi.org/10.1097/ACM.0b013e31822a6cf8
27. Pangaro L, Ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. *Med Teach*. 2013; 35(6):e1197-1210. https://doi.org/10.3109/0142159X.2013.788789
28. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005; 80 Supp: S84-87. https://doi.org/10.1097/00001888-200510001-00023
29. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ.* 2006; 6:42. https://doi.org/10.1186/1472-6920-6-42
30. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: a review of metrics AMEE guide no. 49. *Med Teach*. 2010; 32(10): 802-811. https://doi.org/10.3109/0142159X.2010.507716
31. Osler W. Examinations, examiners and examinees. *Lanc*. 1913; 182(4705): 1047–1050. https://doi.org/10.1007/BF02964451
32. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa surgical competency operating room evaluation (O-SCORE): A tool to assess surgical competence. *Acad Med.* 2012;87(10);1401-7. https://doi.org/10.1097/ACM.0b013e3182677805
33. Colliver JA. Validation of standardized patient assessment: a meaning for clinical competence. *Acad Med.* 1995; 70 (12): 1062- 1064. https://doi.org/10.1097/00001888-199512000-00006
34. Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of longitudinal data.* Oxford Statistical Science Series. 25. 2nd ed. Oxford, New York: Oxford University Press; 2002.
35. Rabe-Hesketh S, Skrondal A. *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press; 2005.
36. Singer JD, Willett JB. *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford: Oxford University Press; 2003.
37. Twisk JWR. *Applied longitudinal data analysis for Epidemiology: A practical guide*. Cambridge: Cambridge University Press; 2003.
38. StataCorp. *Stata Statistical Software: Release 14. College Station*, TX: StataCorp LP; 2015.
39. Holmboe ES, Fiebach NH, Galaty LA, Huot S. Effectiveness of a focused educational intervention on resident evaluations from faculty: A randomized controlled trial. *J Gen Intern Med*. 2001; 16(7): 427-434. https://doi.org/10.1046/j.1525-1497.2001.016007427.x
40. Oller KL, Mai CT, Ledford RJ, O'Brien KE. Faculty development for the evaluation system: a dual agenda. *Adv Med Educ Prac.* 2017; 8: 205-210. https://doi.org/10.2147/AMEP.S124004
41. Murphy MJ, Seneviratne RDA, Remers OJ, Davis MH. "Hawks" and "doves": effect of feedback on grades awarded by supervisors of student selected components. *Med Teach*. 2009; 31(10): e484-e488. https://doi.org/10.4300/JGME-D-14-00161.1
42. Raj JM, Thorn PM. A faculty development program to reduce rater error on milestone-based assessments. *J Grad Med Educ.* 2014; 6(4):680-685. https://doi.org/10.4300/JGME-D-14-00161.1
43. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009; 24(1): 74-79. https://doi.org/10.1007/s11606-008-0842-3
44. Tavares W, Eva, KW. Exploring the impact of mental workload on rater-based assessments. *Adv Health Sci Educ Theory Pract*. 2013; 18(2): 291-303. https://doi.org/10.1007/s10459-012-9370-3
45. Choi H-H, van Merriënboer JG, Paas F. Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educ Psych Review*.

2014; 26(2): 225-244. https://doi.org/10.1007/s10648-014-9262-6

46. Zhao R, D'Eon M. Five ways to get a grip on grouped self-assessments of competence for program evaluation. *Can Med Educ J*. 2020; 11(4): e90-e96. https://doi.org/10.36834/cmej.69276
47. Watson, RM., Nuhfer W, Moon KN, et al. Paired measures of competence and confidence illuminate impacts of privilege on college students. *Num*. 2019; 12(2). https://doi.org/10.5038/1936-4660.12.2.2
48. D'Eon MF, Trinder K. Evidence for the validity of grouped self-assessments in measuring the outcomes of educational programs. *Eval Health Prof.* 2014; 37(4):457-69. https://doi.org/10.1177/0163278713475868