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, Danika Vautour, Michael McMullen, Nicholas Cofie, Nancy Dalgarno, Rachel Phelan et Glenio B Mizubuti

Volume 12, numéro 4, 2021
URI : https://id.erudit.org/iderudit/1082647ar
DOI : https://doi.org/10.36834/cmej.70697

Résumé de l'article
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 globale par 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 (xi(enseignant) - xi(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.
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 Mizubuti2
1Department of Anesthesiology and Perioperative Medicine, Kingston Health Sciences Centre; 2Faculty 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; published: September 14, 2021; CMEI 2021, 12(4). Available at http://www.cmej.ca
© 2021 Fleming, Vautour, Cofie, McMullen, Dalgarno, Phelan, Mizubuti; licensee Synergies Partners
https://doi.org/10.36834/cmej.70697. This is an Open Journal Systems article distributed under the terms of the Creative Commons Attribution License. (https://creativecommons.org/licenses/by-nc-nd/4.0) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited.

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_{assessor}−X_{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 (r_\text{ho} = 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_{enseignant}−X_{resident}), 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 (r_\text{ho} = 0.70) que ne l’étaient les enseignants. Les scores des 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-confidence 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. It reflects one’s ability to make judgments regarding their own competence, and is critical for learners’ self-regulation, and professional development. As a skill, self-assessment requires deliberate training and practice. 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.

Evidence from the medical literature suggests that residents generally do poorly at self-assessing their performance as compared to faculty evaluation. 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. 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. 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. 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.

The accuracy of learners’ assessments is affected by rater bias and/or inaccuracies which may be systematic or non-systematic. 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. 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. Assessor bias is an important source of rater error in faculty assessment of residents, and reduces the reliability and validity of such assessments. Rater errors include those arising from rater variance, extreme rater stringency, and leniency biases. 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. McManus, Thompson and Mollon, 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. and Bartmann et al. 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.

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, 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, 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. 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**

<table>
<thead>
<tr>
<th>Rating</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“I had to do it completely”: Requires complete hands on guidance, did not do, or not given the opportunity to do.</td>
</tr>
<tr>
<td>2</td>
<td>“I had to intervene or talk the resident through”: Able to perform task but needs constant direction.</td>
</tr>
<tr>
<td>3</td>
<td>“I had to prompt them from time to time”: Demonstrates some independence but requires intermittent direction.</td>
</tr>
<tr>
<td>4</td>
<td>“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.</td>
</tr>
<tr>
<td>5</td>
<td>“I did not need to be there”: Completely independent, understands risks and performs safely, insightful, pre-emptive and proactive, ready for practice.</td>
</tr>
</tbody>
</table>

**Outcome variable**

The main outcome variable of interest is accuracy of self-assessments. Following Gordon’s 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. Thus, using faculty scores as the gold standard, we compared faculty scores with residents’ scores \(X_{i\text{faculty}} - X_{i\text{resident}}\) 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 \(X_{i\text{faculty}} - X_{i\text{resident}} < 0\) or under-confident \(X_{i\text{faculty}} - X_{i\text{resident}} > 0\) 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. This procedure allows the regression parameter estimates to be unbiased and efficient. Stata statistical software was used to perform the analysis.

**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%).
Table 2 shows that the majority of accurate faculty scores, averaging 23.0% of all of their assessments. Most residents (86.7%) scored 1.0 in excess of faculty assesso
13.3% of the residents scored 1.5 less and 1.5 more than these particular self-assessments. 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

<table>
<thead>
<tr>
<th>Faculty assessor type</th>
<th># of assessments</th>
<th>Mean /SD</th>
<th>Range of scores</th>
<th># of Scores &lt; 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stringent</td>
<td>2</td>
<td>3.05</td>
<td>(1.15)</td>
<td>2 - 5</td>
</tr>
<tr>
<td>Neutral</td>
<td>4</td>
<td>3.8</td>
<td>(0.65)</td>
<td>2 - 5</td>
</tr>
<tr>
<td>lenient</td>
<td>1</td>
<td>4.67</td>
<td>(0.26)</td>
<td>4.5 - 5</td>
</tr>
</tbody>
</table>

Note: Levene’s robust test statistic (Wo) 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

<table>
<thead>
<tr>
<th>Faculty assessor type</th>
<th>Over-confidence scores</th>
<th>Accurate scores</th>
<th>Under-confidence scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stringent</td>
<td>27.78</td>
<td>2.63</td>
<td>3.91</td>
</tr>
<tr>
<td>Neutral</td>
<td>72.22</td>
<td>97.37</td>
<td>92.74</td>
</tr>
<tr>
<td>lenient</td>
<td>0.00</td>
<td>0.00</td>
<td>3.35</td>
</tr>
</tbody>
</table>

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

Note: Levene’s robust test statistic (Wo) for the equality of variances between the assessor groups = 14.203, df(2, 236), Prob > F =0.00000.
assessment scores were on average, 0.4 points higher ($\alpha = 0.396, z = 4.39, p < 0.001$) 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 ($\theta_2 = 0.40, CI = 0.376 – 0.433$) was much lower than the within faculty variance observed ($\theta_2 = 0.64, CI = 0.598 – 0.692$). On the other hand, the between faculty assessment scores appear to be more similar with less variability ($\psi^2 = 0.32, CI = 0.236 – 0.444$) compared to the between resident self-assessment scores ($\psi^2 = 0.61, CI = 0.423 – 0.882$).

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

| Coef. | Std. err | z     | p > |z|  | 95% Confidence interval |
|-------|----------|-------|------|-----|------------------------|
| 0.396 | 4.39     | 0.00  | 0.219| 0.572         |

Random intercepts
- Between residents variance ($\psi^2$)
- Within residents variance ($\theta^2$)
- Rho (intra-class correlation)

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

<table>
<thead>
<tr>
<th>Coef.</th>
<th>Std. err</th>
<th>z</th>
<th>p &gt;</th>
<th>z</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3.58</td>
<td>0.16</td>
<td>22.39</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Random intercepts
- Between residents variance ($\psi^2$)
- Within residents variance ($\theta^2$)
- Rho (intra-class correlation)

Faculty scores
- Constant ($\alpha$)

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 ($\beta = 0.121, z = 3.16, p < 0.01$) or neutral faculty assessors ($\beta = 0.137, z = 3.57, p < 0.001$). Senior residents were significantly less likely to be under-confident compared with junior residents ($\beta = -0.182, z = -2.45, p < 0.05$). The year of assessment had no effect on residents’ accuracy ($\beta = 0.047, z = 0.58, p > 0.05$).

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, 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.

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, although such interventions could still result in faculty scores that are more lenient or more stringent. 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. Yet, others point out that this strategy might be less effective since medical raters may be unresponsive to

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

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Std. err</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>∆(Faculty-resident scores)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant (α)</td>
<td>0.45***</td>
<td>0.08</td>
<td>5.48</td>
</tr>
<tr>
<td>Faculty assessor type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stringent (Ref. category)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>0.14***</td>
<td>0.04</td>
<td>3.57</td>
</tr>
<tr>
<td>Lenient</td>
<td>0.12**</td>
<td>0.04</td>
<td>3.16</td>
</tr>
<tr>
<td>Resident Seniority</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior (Ref. category)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior</td>
<td>-0.18*</td>
<td>0.07</td>
<td>-2.45</td>
</tr>
<tr>
<td>Year of assessment</td>
<td>0.05</td>
<td>0.08</td>
<td>0.58</td>
</tr>
<tr>
<td>Period of assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First quarter (Ref. category)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second quarter</td>
<td>0.01</td>
<td>0.15</td>
<td>0.13</td>
</tr>
<tr>
<td>Third quarter</td>
<td>0.08</td>
<td>0.14</td>
<td>0.58</td>
</tr>
<tr>
<td>Random intercepts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between residents variance</td>
<td>0.27</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>(ψ)²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within residents variance</td>
<td>0.60</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>(α)²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rho (intra-class correlation)</td>
<td>0.16</td>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Faculty assessor type</th>
<th>Stringent</th>
<th>Neutral</th>
<th>Lenient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy of assessments</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Over-confidence scores</td>
<td>50.00</td>
<td>8.58</td>
<td>0.00</td>
</tr>
<tr>
<td>Accurate scores</td>
<td>15.00</td>
<td>36.63</td>
<td>0.00</td>
</tr>
<tr>
<td>Under-confidence scores</td>
<td>35.00</td>
<td>54.79</td>
<td>100.00</td>
</tr>
<tr>
<td>N</td>
<td>20</td>
<td>303</td>
<td>6</td>
</tr>
</tbody>
</table>

...
professional development (e.g., Cook et al.). 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. 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 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.}\textsuperscript{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**

https://doi.org/10.1097/00001888-200510001-00015


35. Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modeling using Stata. College Station, TX: Stata Press; 2005.


