Comparing 2 Approaches for the File Review of Residency Applications

Nada Gawad, MD, MAEd
Julia Younan, MD
Chelsea Towaij, MD
Isabelle Raiche, MD, MAEd, FRCSC

ABSTRACT

Background The residency selection process relies on subjective information in applications, as well as subjective assessment of applications by reviewers. This inherent subjectivity makes residency selection prone to poor reliability between those reviewing files.

Objectives We compared the interrater reliability of 2 assessment tools during file review: one rating applicant traits (ie, leadership, communication) and the other using a global rating of application elements (ie, curriculum vitae, reference letters).

Methods Ten file reviewers were randomized into 2 groups, and each scored 7 general surgery applications from the 2019–2020 cycle. The first group used an element-based (EB) scoring tool, while the second group used a trait-based (TB) scoring tool. Feedback was collected, discrimination capacities were measured using variation in scores, and interrater reliability (IRR) was calculated using intraclass correlation (ICC) in a 2-way random effects model.

Results Both tools identified the same top-ranked and bottom-ranked applicants; however, discrepancies were noted for middle-ranked applicants. The score range for the 5 middle-ranked applicants was greater with the TB tool (6.43 vs 3.80), which also demonstrated fewer tie scores. The IRR for TB scoring was superior to EB scoring (ICC [2, 5] = 0.82 vs 0.55). The TB tool required only 2 raters to achieve an ICC $\geq$ 0.70.

Conclusions Using a TB file review strategy can facilitate file review with improved reliability compared to EB, and a greater spread of candidate scores. TB file review potentially offers programs a feasible way to optimize and reflect their institution's core values in the process.

Introduction

The residency selection process, which generally consists of both a file review and candidate interview,[1] is a difficult and subjective task. Despite the goal of both components to ultimately identify candidates who would perform best in a given program, literature has demonstrated that few elements of a traditional application process predict clinical performance during and after residency.[2,3] Furthermore, the selection process is resource-intensive and consumes a significant number of faculty hours.[3] The Canadian Resident Matching Service (CaRMS) process takes place from November to March, beginning with applicant file review and distribution of interview invitations, followed by a nationwide 2-week interview period of invited applicants, and subsequent ranking and matching. Using only objective data for the process would certainly be less time consuming, and studies have demonstrated correlation between pre-residency examination scores and in-training examination scores.[4] However, in-training examination scores are only one metric of clinical performance, and using only objective data has been shown to be a poor predictor of clinical performance overall.[5] Additionally, in Canada, grades and national examination scores are provided as pass or fail and are thus of little value in the file review process. Applicant traits found in the curriculum vitae (CV) and personal letter specifically have shown correlation with clinical performance and decreased resident attrition in a few studies,[6,7] suggesting further investigation is warranted.

Scoring the subjective components of the applicant file (ie, personal letter, letters of reference, CV) can be challenging and is prone to poor reliability between raters, particularly with respect to subjective data.[8] A prior study investigating the interrater reliability in scoring individual elements of the applicant file, such as the personal statement, found it to be highly subjective with significant variability in scores.[9] Specifically, a lack of objective criteria for evaluation of the subjective measure led to contradictory evaluations and inconsistent rationale between raters.[9] Suggestions for improving the reliability of the file review scoring system include using specific

DOI: http://dx.doi.org/10.4300/JGME-D-20-00619.1

Editor's Note: The online version of this article contains the 2 scoring tools used in the study.

applicant traits valued by the program, as opposed to a global rating.[2,8]

In our own program, we questioned the interrater reliability within our traditional application review process in which we applied a global rating score to the subjective application file elements (ie, CV, reference letters, personal statement, etc). Specifically, we noted clustering of scores for our middle-ranked candidates,[10] making it difficult to discern which of these candidates should be offered our limited interview slots. Therefore, the objective of this study was to compare 2 approaches to file review: our historic review process focusing on a global assessment of file elements (ie, CV, reference letters, personal statement, etc) and our proposed new file review process focusing on scoring specific applicant traits valued by our program (ie, leadership, communication, compassion, etc) and found within the application elements to determine which offered the best interrater reliability and greater spread of applicant scores.

Objectives
To compare the interrater reliability of 2 assessment tools during file review: one rating applicant traits and the other using a global rating of application elements.

Findings
The trait-based tool demonstrated a superior interrater reliability, with a greater range of scores for middle-ranked applicants.

Limitations
In addition to a limited sample size, further research is needed to assess if trait-based file review leads to selection of better residents and translatability of the process in this study to other programs.

Bottom Line
Using a trait-based file review strategy can offer programs a feasibly way to reflect their institutions' core values in the selection process with improved reliability and a greater spread of middle-ranked applicant scores.

Methods

To facilitate the design of a new "trait-based" (TB) application scoring tool, an informal survey was sent to the University of Ottawa Division of General Surgery staff and resident surgeons to determine what applicant traits were deemed important to success as a resident and a career in surgery, and which should be avoided. A TB scoring tool was then created by the residency application committee leads (N.G. and I.R.) based on the highest ranked traits (provided as online supplementary data). This new tool scored files based on traits such as teamwork, leadership, perseverance, compassion, and teachability, and provided examples of what characteristics would merit each score. The previously used scoring tool was labeled "element-based" (EB) because it provided a global rating score for each element of the applicant file (ie, curriculum vitae, reference letters, personal statement; provided as online supplementary data).

To compare the 2 scoring tools, 10 members of the residency application committee (4 staff surgeons, 6 surgical residents) were randomly assigned to a scoring tool with equal distribution of staff and residents in each group (TB versus EB), representative of reviewer expertise in our file review process wherein there are teams of reviewers made up of staff and residents (postgraduate year 2 and above). To ensure sufficient experience and consistency, we grouped new reviewers with veteran reviewers. In December 2018, just prior to the CaRMS file review process, each reviewer independently and prospectively reviewed 7 applicant files using their assigned tool sent via email. The reviewers had all previously performed file review. The training for both scales was minimal and included advising the raters to use the full extent of the scoring scale and to base their assessment only on the applicant file, not on any personal experience they may have had with the applicant.

With feasibility in mind, 7 files were chosen as a purposeful sample, which is consistent with similar research on file review process reliability.[9] The applicant files selected for review were randomly chosen in alphabetical order from the 2019 CaRMS Canadian Medical Graduate applicant pool. International medical graduates were excluded from the study as their applications tend to be more heterogeneous with a wider range of backgrounds and circumstances. The applicant files consisted of medical students from our local institution, other institutions within the province, interprovincial, and American institutions. Some applicants had completed electives at our institution while others had not, and therefore applicants may or may not have interacted with reviewers. This is consistent with our historical CaRMS file review process. The research team audited the files to ensure the sample included applicants with a wide variety of demographics and experiences who were likely to have a wide range of scores. An appropriate level of variety was noted in the initial sample, and thus no changes to the sample were needed.

After the reviewers scored the application files, they were asked to provide open-ended written feedback on the tool. To determine the interrater reliability of the 2 scoring tools (EB and TB), the intraclass correlation (ICC) and 95% confidence intervals were calculated based on an absolute agreement, single-rater, 2-way random-effects model. Interpretation of

TABLE 1
Reviewer Demographics

Demographics	Element-Based	Trait-Based
Gender		
Male	3	2
Female	2	3
Title		
Staff	2	2
PGY-5	1	...
PGY-4	...	...
PGY-3	1	3
PGY-2	1	...

Abbreviation: PGY, postgraduate year.

TABLE 2
Applicant Demographics

Applicant Characteristics	N (%)
Gender	
Male	2 (29)
Female	5 (71)
Qualifications	
CMG	7 (100)
Local medical student	1 (14)
> 50% of electives in general surgery	6 (86)
Interview offers	4 (57)

Abbreviation: CMG, Canadian medical graduate.

the ICC was < 0.5 = poor, 0.5 to 0.75 = moderate, 0.75–0.90 = good, and > 0.90 = excellent interrater reliability.[11] SPSS 25 software (IBM Corp, Chicago, IL) was used for all analyses.

Ethics approval was waived by the Ottawa Health Science Network Research Ethics Board as the primary aim of this study was for quality improvement of the CaRMS selection process.

Results

Each of the 2 scoring tools (TB and EB) was used by 2 staff and 3 resident reviewers. Each reviewer group consisted of both men and women (TABLE 1). Both the TB and EB scoring tools used a possible total of 35 points.

The 7 applicant files consisted of 2 male (29%) and 5 female (71%) applicants, compared to an overall pool of 42% male and 58% female applicants. One medical student was from our local university. The majority (86%) of the applicants completed more than 50% of their electives in general surgery, or subspecialties thereof, including pediatric surgery and thoracic surgery (TABLE 2).

The overall interrater reliability for the applicant files reviewed were ICC (2, 5) = 0.82 (95% CI 0.57–0.96) using the TB tool, compared to 0.55 (95% CI 0.19–0.88) using the EB tool. With respect to reference letters specifically, the TB tool yielded an ICC (2, 5) of 0.82 (95% CI 0.59–0.96) vs 0.30 (95% CI 0.05–0.72) for the EB tool. Curriculum vitae and research productivity demonstrated poor to moderate ICC using both tools (TABLE 3). We additionally calculated the ICC of all possible rater combinations, and found that using the trait-based tool, any 2 raters can achieve an ICC (2, 2) $\geq$ 0.71.

The candidates were ranked from first to seventh based on the mean score from all reviewers. The top- and bottom-ranked candidates were the same with both tools; however, there were discrepancies in rank order within the middle-ranked candidates. The range of scores from the top- to bottom-ranked applicants was 14.14 vs 13.00 (out of 35 points) for the TB versus EB tools. When comparing the middle-ranked candidates (rank 2–6), the score range was 6.43 (TB) versus 3.60 (EB) points (FIGURE). With respect to the scores from each individual reviewer, there were also more tie scores using the EB tool compared to the TB tool (TABLE 4).

Comments collected regarding the new TB tool identified concerns relating to scoring reference letters due to frequent global statements provided by referees (ie, "the best medical student I have worked with"), as opposed to specific trait-based comments or examples. Similar concerns were expressed regarding identification of certain traits within CVs. Additionally, reviewers thought the file review process to be more time consuming with the TB tool, compared to their experience using the EB tool in previous years. Anecdotally, however, we noted that the official file review process took the same number of days compared to prior years, despite the use of the new tool.

Discussion

The results of this study demonstrate that using the TB scoring tool offers better interrater reliability compared to the EB scoring tool, suggesting that it is a more consistently applied method of file review even with a wide range of reviewer experience. It also provides a greater spread of scores for middle-ranked candidates with fewer tie scores, allowing for improved rank list discrimination when determining interview offers.

Certain elements of applicant files are known to be challenging to score, notably the personal statement and reference letters.[8,9] These elements are perceived to be subjective and raters report being uncomfortable attributing a global score to them.[8] Similarly, the EB system asks the rater for a global rating of each element. Authors have voiced concerns regarding the

TABLE 3
Comparison of Interrater Reliability Between Scoring Tools

Scoring Tool	Research	Reference Letters	CV	Total Score
Element-based ICC (2,5) (CI)	0.33 (0.06–0.76)	0.30 (0.05–0.72)	-0.01 (-0.09-0.30)	0.55 (0.19–0.88)
Trait-based ICC (2,5) (CI)	0.41 (0.11–0.80)	0.82 (0.59–0.96)	0.29 (0.05-0.72)	0.82 (0.57–0.96)

Abbreviation: ICC, intraclass correlation.

use of global rating in assessment of certain constructs because the personal characteristics of the assessors are more likely to influence the rating.[12,13] Checklists, or more directed assessment tools, provide more guidance to the rater and are less reliant on rater expertise.[14] Recent scholarly work on global rating scales versus checklist-based rating scales demonstrates similar interrater reliability and discrimination.[14] However, it has been suggested that the choice of a rating scale should depend on the context in which it is to be used.[14] In the context of resident selection, a task that is relatively unfamiliar to physicians, it is possible that the extra guidance offered by the checklist is beneficial. Interestingly, our results show that the highest increase in interrater reliability was found in the assessment of reference letters, one of the most subjective elements of the application. This suggests that the extra guidance led the assessor to give a more uniform score.

Kelz et al[7] demonstrated that the use of a trait-based selection process led to decreased attrition rate in a surgical residency program. They developed an individualized assessment method for each applicant with the help of an organizational management expert. Our study adds to these results by demonstrating a feasible trait-based selection approach. The process of creating the TB assessment tool is simple and straightforward and could allow other programs to create a file review assessment tool that reflects their values. Furthermore, by showing that only 2 reviewers are required to achieve an ICC $\geq$ 0.71, our study suggests that programs could implement a trait-based assessment tool without having to increase the number of raters required for reliable assessment, which improves file review efficiency despite initial reviewer feedback that the TB tool was more time-consuming.

Feedback from raters noted some difficulty identifying specific traits within the applicant files, resulting in a perception of artificially low scores compared with the global impression suggested by specific positive descriptors in the reference letters (eg, "in the top 5% of medical students," "functioning at a resident level"). This perception of the potential risks of missing competency has been discussed with the use of a more guided scale by other authors.[14] The rater concerns are interesting in light of a study which showed that the number of times an Accreditation Council for Graduate Medical Education competency was named in a reference letter correlated to residency success.[5] Accordingly, the results of the latter study suggest it is appropriate to give points for the presence of specific details without concern of disadvantaging applicants

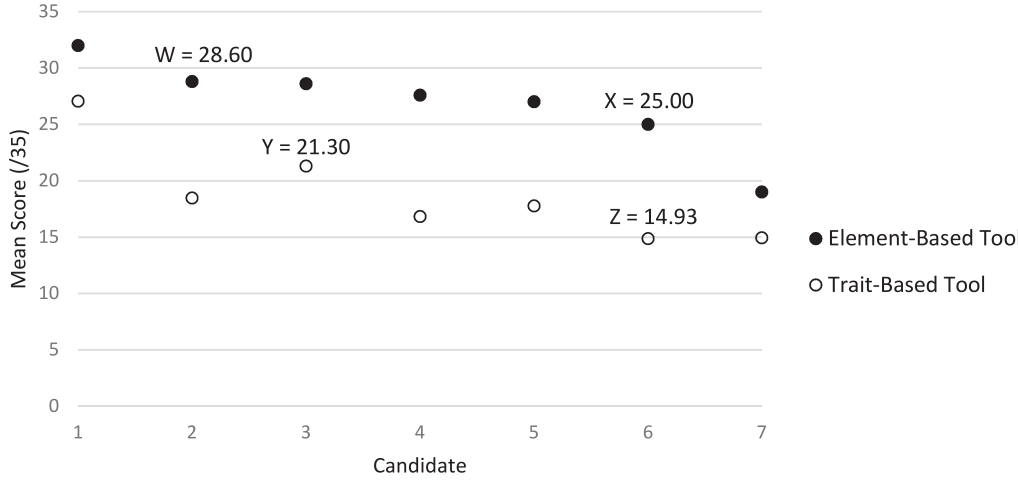

FIGURE
Comparison of Mean Overall Candidate Scores
Note: Score ranges calculated for mid-ranked candidates (ΔW-X = 3.60 vs ΔY-Z = 6.43).

TABLE 4
Number of Tie Candidate Scores Per Rater

Rater	Element-Based Tool
1	2
2	1
3	1
4	1
5	1[a]
Rater	**Trait-Based Tool**
1	1
2	0
3	1
4	0
5	0

[a] Denotes a 3-way tie of candidate scores.

with less detailed letters. Further study is needed to ensure fair assessment of the trainee.

Our study has limitations. The sample size was limited to 7 files, due to time constraints relating to CaRMS file availability as well as rater availability, as the files had to be reviewed between the time they were released by CaRMS and with sufficient time to complete the official CaRMS file review thereafter. This is, however, consistent with similar literature.[8] Also, the applicant sample used was predominantly female, with one additional female resident relative to the proportion of females in the applicant pool. While a male applicant could have been substituted in the prescreening process, for a difference of just one applicant we thought the risk of introducing bias in the interest of ensuring a perfectly representative sample of sufficient variety was not warranted. Finally, ideally our study would have then compared the file review ranking with each rating scale to interview performance. However, of the applicant pool studied, only 4 of 7 applicants were offered interviews, making it difficult to make any meaningful observations on comparison with interview performance.

Further research is needed to assess if trait-based file review leads to selection of better residents and the extent to which the process used in this study to develop and implement a trait-based review process is translatable to other programs.

Conclusions

Using a trait-based file review strategy can facilitate file review with better reliability and a greater spread of middle-ranked candidates than global element-based scoring. Trait-based file review potentially offers programs a feasible way to both optimize and reflect their institution's core values in the file review process.

References

1. Canadian Resident Matching Service (CaRMS). File review & interviews. https://www.carms.ca/match/r-1-main-residency-match/applicant/file-review-interviews-r1/. Accessed January 15, 2021.
2. Borowitz SM, Saulsbury FT, Wilson WG. Information collected during the residency match process does not predict clinical performance. *Arch Pediatr Adolesc Med*. 2000;154(3):256–260. doi:10.1001/archpedi.154.3.256.
3. Wall J, Votey SR, Solomon T, Schriger DL. Is National Resident Matching Program rank predictive of resident performance or post-graduation achievement? 10 years at one emergency medicine residency. *West J Emerg Med*. 2019;20(4):641–646. doi:10.5811/westjem.2019.4.40602.
4. Swanson DB, Sawhill A, Holtzman KZ, et al. Relationship between performance on part I of the American board of orthopaedic surgery certifying examination and scores on USMLE Steps 1 and 2. *Acad Med*. 2009;84(10 suppl):21–24. doi:10.1097/ACM.0b013e3181b37fd2.
5. Stohl HE, Hueppchen NA, Bienstock JL. Can medical school performance predict residency performance? Resident selection and predictors of successful performance in obstetrics and gynecology. *J Grad Med Educ*. 2010;2(3):322–326. doi:10.4300/jgme-d-09-00101.1.
6. Naylor RA, Reisch JS, Valentine RJ. Factors related to attrition in surgery residency based on application data. *Arch Surg*. 2008;143(7):647–651. doi:10.1001/archsurg.143.7.647.
7. Kelz RR, Mullen JL, Kaiser LR, et al. Prevention of surgical resident attrition by a novel selection strategy. *Ann Surg*. 2010;252(3):537–541. doi:10.1097/SLA.0b013e3181f27a50.
8. Dirschl DR. Scoring of orthopaedic residency applicants: is a scoring system reliable? *Clin Orthop Relat Res*. 2002;(399):260–264. doi:10.1097/00003086-200206000-00033.
9. White BAA, Sadoski M, Thomas S, Shabahang M. Is the evaluation of the personal statement a reliable component of the general surgery residency application? *J Surg Educ*. 2012;69(3):340–343. doi:10.1016/j.jsurg.2011.12.003.
10. Karan S. Confessions of a Program Director—Making My Program's Rank List. *Thalamus*. https://thalamusgme.com/confessions-of-a-program-director-making-my-programs-rank-list/. Accessed January 15, 2021.
11. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012.

12. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' slinical skills? The impact of faculty's own clinical skills. *Acad Med*. 2010;85(10 suppl):25–28. doi:10.1097/ACM.0b013e3181ed1aa3.
13. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Heal Sci Educ*. 2013;18(3):325–341. doi:10.1007/s10459-012-9372-1.
14. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ*. 2015;49(2):161–173. doi:10.1111/medu.12621.

All authors are with the University of Ottawa, Ottawa, ON, Canada. **Nada Gawad, MD, MAEd,** is a Resident, Department of Surgery; **Julia Younan, MD,** is a Resident, Department of Surgery; **Chelsea Towaij, MD,** is a Resident, Department of Surgery; and **Isabelle Raiche, MD, MAEd, FRCSC,** is Assistant Professor of Surgery, and Ottawa General Surgery CaRMS Director, Department of Surgery.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

Corresponding author: Nada Gawad, MD, MAEd, University of Ottawa, Ottawa, ON, Canada, ngawad@toh.ca, Twitter @ngawadMD

Received June 11, 2020; revisions received October 13, 2020, and December 11, 2020; accepted December 15, 2020.