The Journal of Education in Perioperative Medicine

Original Research

Quality Control for Residency Applicant Scores

Jed Wolpaw, MD, MEd
Gillian Isaac, MD, PhD
Tina Tran, MD
Mike Banks, MD
Steven Beaudry, DO
Priyanka Dwivedi, MA
Serkan Toy, PhD

Introduction

Most residency programs in the United States use a candidate selection process. The intended utility of this process is to ensure selection of the most qualified candidates who are the best fit and therefore most likely to succeed. This process involves a diverse group of faculty judges/interviewers making inferences based on, typically, a review of application materials, interviews, and faculty group discussions. The quality of the resulting rank list depends on measurement precision and accuracy. A precise measurement model produces linear and reproducible measures, and accurate measurement allows for targeting the actual candidate's ability, free from confounders such as different faculty interviewers on different days.[1] Indeed, interrater reliability is low in interview scoring.[2, 3]

We were unable to identify any commonly used quality control methods for evaluating the scores from interviewers before a rank list is made. And yet, the potential for poor quality data is real. A study found faculty interviewing candidates for medical school differed significantly in their degree of stringency or leniency.[4] Using different interviewers on different days has the potential to inflate or deflate scores for candidates on one day compared to candidates on another day.

Many-facet Rasch measurement (MFRM) is a family of measurement models that allow for establishing a quality control system for rater-mediated assessment that can identify these outlier scores. This model has been proven useful for quality control in undergraduate medical education admissions.[4-6] We hypothesized that using an MFRM model, we could establish a quality control system to identify noise in our score data and address potential sources of measurement error to produce fair averages for each candidate.

Methods

This is an observational study that took place at a large academic medical center from October 2017 to January 2018. The local institutional review board deemed this to be a quality improvement project. The department in which this study took place interviews 160 candidates each year for 25 available spots. All interviews are conducted by 4 faculty members who interview 8 candidates per day. Two faculty members—the program director and associate program director—interview all 160 candidates. The third interviewer spot is taken up by 1 of 3 assistant program directors. The fourth interviewer is a faculty member who signs up to interview. Because the fourth interviewer was almost always a different faculty member (17 different faculty members filled this spot over the course of 20 interview days), and thus only gave 1 set of scores. We had to exclude their scores from our analysis in order to establish connectivity in the dataset. Interviewers are given a description of the scoring scale, which ranged from 1 to 100. The scale has been used by our department for many years and defines a score in the 90s as someone who could be a chief resident, a score in the 80s as someone who we would be happy to have, a score in the 70s as someone who would probably do fine, a score in the 60s as someone we might put at the bottom of our rank list, and a score below 60 as someone we would consider not ranking at all. This scale is sent to each interviewer along with the applications to read. Interviewers gave 3 sets of scores, which were entered into Qualtrix (Provo, Utah). The first score was given after reading the application but before the interview. The second score was given after the interview but before the group discussion in which all 4 interviewers discussed the candidates. The final score was given after the discussion.

Data Analysis

Using MFRM, one can examine multiple variables (facets) that might be potential sources of variance for the outcome variable.[7] For example, in addition to examinees' ability (candidate fitness for the residency program), testing situation (or scoring occasion, ie review of application documents vs. interview), or rater leniency/severity could also influence the scores candidates receive. This psychometric approach produces standardized indices for determining the degree to which the data fit the expectations predicted by the model. Expected fair averages for candidate qualification/fit are calculated based on the observed values adjusted for rater leniency/severity and/or task difficulty (scoring occasion). The difference between the observed and expected values, called standardized residuals, indicates the quality of the data and the accuracy of the measurement.[1]

There are 2 types of mean-square (MnSq) fit indices, outfit and infit, that help flag the values presenting misfit. Outfit MnSq

continued on next page

continued from previous page

is the unweighted mean of the squared standardized residuals and is sensitive to outliers. Infit MnSq values are the weighted mean squared residual goodness of fit statistics. Infit MnSq indices are not affected by the outliers and could have a better utility for diagnosing misfitting examinees/candidates, judges, or tasks.[8] For both statistics, an expected MnSq value is 1 when the model fits the data, and can vary from 0 to infinity.[9] For rating scales, values greater than 2 might reveal severe misfits (underfit, unmodeled noise) that could be distorting or degrading the measurement model. On the other hand, values lower than 0.5 indicate too little variation (overfit), which could indicate that ratings produce redundant information or that there is a restriction of range in the use of the scale. Conventionally, infit MnSq values within the 0.5 to 1.5 range indicate a model useful for constructing measurement.[10] Z-standardized (*Zstd*) statistic further indicates statistical significance clarifying whether or not the misfitting values are occurring by chance.

MFRM analysis also produces estimates for examining the differences in judge leniency/severity. A fixed χ^2 statistic tests the hypothesis that the judge leniency/severity measures are not significantly different. Failing to reject the null hypothesis ($P < .05$) means that at least 2 of the judges' leniency/severity measures differed significantly.[11]

In the present study, we used a free version of the FACETS software, MINIFAC, for the analysis (Winsteps.com, Beaverton, OR). This program does not require that every judge rate every candidate in all occasions. As long as observations are connected through a network linking every parameter for all facets, estimates for judges, candidates, and occasions can be obtained for each of these facets independent of the other facets, which can then be calibrated together on a common logit scale.[12] This method allows for an examination of the extent to which specific elements within each facet (ie, individual faculty judges) contribute to noise in the measurement system. Our model included 3 facets: faculty interviewers, candidates, and occasions.

Results

A total of 1378 observations were used in the MFRM model, explaining 58.42% of the variance in the data. The dataset contained connectivity, allowing the software to produce estimates for each of the elements within every facet (judges, candidates, and occasions), calibrate them on a common logit scale, and place them on a frame of reference, known as "Wright map." Figure 1 shows all 3 facets on a logit scale (left column), with candidate qualifications/fit and occasion easiness centered on the mean of zero logits and faculty judges/raters allowed to float. All 3 facets are positively oriented, meaning that higher logit scores indicate more lenient judges, better qualified/fit candidates, and easier scoring occasions (whether candidates received consistently higher scores at any of the 3 time points: application review, interview, or interviewer group discussion).

The Wright map in Figure 1 indicates that, overall, all 5 raters (second column) had positive logit scores—they were mostly lenient. Judge 1 seems to be the most lenient (2.02 logits) and Judge 4 the least (1.84 logits). The other 3 judges had about the same level of leniency. Candidates are shown in the third column; each asterisk indicates 2 candidates. Candidate qualification/fit measures ranged from –1 logits to 1.46 logits. The occasions were all clustered at the average logit value and did not seem to vary in terms of easiness/difficulty (scores assigned to candidates at 3 time points were similar).

We also examined whether the judges differed significantly in their leniency and whether any judge's ratings showed misfit to the model (values < 0.05 and > 1.5). The χ^2value of 56.2 with 4 degrees of freedom indicated that at least 2 of the judges' leniency measures differed significantly $P < .001$). To diagnose misfit, we examined the fit indices for judges. Table 1 shows summary statistics for faculty judges, including the total number of assigned ratings (3 occasions) and their fit indices. The infit MnSq value of 2.02 for Judge 3 indicates inconsistent application of the rating scale. The infit *Zstd* (6.8) shows that this was not likely to occur by chance. All other judges had goodness-of-fit measures within the acceptable range for constructing measurement.

Many-facet Rasch analysis output included some unexpected observations based on standardized residual values of 3 and above. Table 2 shows a summary of the unexpected observations. This information can help program directors focus on specific instances for diagnosing measurement error. We will highlight some examples.

First is a candidate (118 in Table 2) who received 4 high scores after the application was read. After the interview, 3 of the 4 interviewers kept their scores high, but the fourth interviewer dropped her score by 20%. This candidate was ranked too low to match but might have been able to match without this low score bringing his net score down. When asked about it in retrospect, the interviewer who gave the low score said that she might have been in a bad mood that day and took offense at some remarks that she normally would not have minded.

Second is a candidate (17 in Table 2) who received mediocre scores from 3 of the interviewers and an extremely high score from a fourth. This candidate was ranked higher than he would have been if not for this outlier high score. When we identified this with our statistical model we looked back and realized that this interviewer was interviewing for the first time that day and had given higher scores that day than on any subsequent day throughout the interview season.

Number 43 in Table 2 received low scores from 3 interviewers and a relatively high score from a fourth. This turned out to be a mistake. When the interviewer looked back at his notes, he had meant to record a lower score than what had been entered in the computer system.

Similarly, number 24 in Table 2 turned out to be a mistake. The interviewer who gave the unusually low score had mixed up part of the printed out application from another candidate and this candidate and therefore gave a mistakenly low score after reading the application.

Discussion

The results showed that this quality control system could be used during the residency interview process to identify misfitting scores that could be further examined. The fact that a score is an outlier does not

continued on next page

continued from previous page

mean it necessarily should be discounted. For example, if a candidate made a racist or sexist comment to one interviewer and not to the others, it would be very reasonable for the one interviewer to give a much lower score than the others after the interview and for the group to then decide to rank that candidate much lower (or not at all).

However, interview scoring has poor interrater reliability, and interviewer scores can be influenced by many factors, including whether or not the reviewers see board scores, whether they see the student's academic record, and the perceived connection made with the interviewee.[2,4] Indeed, variance in interview scores has been shown to be caused in part by the stringency or leniency of individual interviewers.[4] It is therefore likely, as in our examples above, that at times, some outliers will be caused by factors related to the interviewer rather than the interviewee. Program directors may want a way to identify these occurrences and choose to adjust candidate scores before making a rank list.

Our analysis identified 2 ways in which we believe we can improve on our scoring and ranking process in the future. First, we currently have only 1 score that each interviewer gives based on their overall impression of the candidate. Some interviewers may place more weight on the interview and others may place more on the application itself. Consequently, a bad interview can have an enormous effect on any 1 score. In the future we plan to compartmentalize scores so that each category (of which the interview will be one) will count for only 20% of the overall score. This change will also allow us to see if outliers are related more to the interview or to another category of scoring.

Second, this process showed us that we can identify outliers and that some will be caused by issues that should not unduly penalize (or benefit) any given candidate. We will plan to run this analysis again next year, but we will do it before, not after, making the rank list. We will then use the information from this analysis to revisit outlier scores and decide whether to adjust them before making a rank list.

This study had some limitations. First, this process requires extra work for programs that are already extremely busy during interview season. Second, the fact that we did not use a rubric for scoring may have led to more outlying scores than would have been seen by programs that do use an analytical rubric.

Quality assurance is a must for any high-stakes process, and this is no exception. The process we describe here is a relatively low-cost, efficient way to test the quality of the data. Future studies should investigate whether outliers are more likely in certain categories of scoring, such as the interview or the application itself.

References

1. Wright BD, Mok MM. An overview of the family of Rasch measurement models, In: Smith EV, Smith RM (Eds.). *Introduction to Rasch measurement: theory, models and applications.* Maple Grove, MN, JAM Press. 2004:1-24.
2. Stephenson-Famy A, Houmard B, Oberoi S, et al. Use of the interview in resident candidate selection: a review of the literature. *J Grad Med Educ*. 2015;7(4):539-48.
3. Smilen SW, Edmund FF, Bianco AT. Residency selection: Should interviewers be given applicants' board scores? *Am J Obstet Gynecol.* 2001;184(3):508-13.
4. Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Med Educ*. 2010;44(7):690-8.
5. Till H, Myford C, Dowell J. Improving student selection using multiple mini-interviews with multifaceted Rasch modeling. *Acad Med*. 2013;88(2):216-23.
6. Sebok SS, Luu K, Klinger DA. Psychometric properties of the multiple mini-interview used for medical admissions: findings from generalizability and Rasch analyses. *Adv Health Sci Educ Theory Pract*. 2014;19(1):71-84.
7. Eckes T. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*, 2nd ed. Peter Lang, Bern, Switzerland, 2015.
8. Wright B, Masters G. *Rating Scale Analysis: Rasch Measurement*, MESA Press, Chicago, IL, 1982.
9. Linacre JM. What do infit and outfit, mean-square and standardized mean? *Rasch Meas Trans*. 2002;16(2):878.
10. Linacre, JM. Size vs. significance: Standardized chi-square fit statistic. *Rasch Meas Trans*. 2003;17(1): 918.
11. Myford C, Wolfe E. Detecting and measuring rater effects using many-facet Rasch measurement: Part I, In: Smith EV, Smith RM, eds. *Introduction to Rasch Measurement*: *Theory, Models and Applications*, Maple Grove, MN, JAM Press. 2004:460-515.
12. Linacre JM, Wright B. Construction of Measures from Many-Facet Data, In: Smith EV, Smith RM, eds. *Introduction to Rasch measurement: Theory, models and applications*, Maple Grove, MN, JAM Press. 2004:296-321.

The authors are at Johns Hopkins Department of Anesthesiology and Critical Care Medicine in Baltimore, MD. ***Jed T. Wolpaw*** *is an Assistant Professor and Residency Program Director;* ***Gillian Isaac*** *is an Assistant Professor;*

Tina Tran *is an Assistant Professor;* ***Michael Banks*** *is an Assistant Professor;* ***Steven Beaudry*** *is an Assistant Professor;* ***Priyanka Dwivedi*** *is a Residency Program Manager;* ***Serkan Toy*** *is an Assistant Professor.*

The authors have no relevant funding and no conflicts of interest. This manuscript has never been published and is not currently under consideration for publication with any other journal.

Abstract

Background: The residency program selection process incorporates application review and candidate interviews to create an ordered rank list. Though this is the single most important process for determining the department's future trainees, the system lacks a quality control mechanism by which faculty ratings are scrutinized. This study used many-facet Rasch measurement (MFRM) to establish a quality control system for the candidate selection process.

Methods: This study took place from October 2017 to January 2018 at a large anesthesiology residency program with 25 available spots. Every candidate received scores from 3 faculty judges across 3 occasions: application review, interview, and interviewer group discussion. MFRM with 3 facets—faculty judges, candidates, and occasions—was used to identify sources of measurement error and produce fair averages for each candidate.

Results: A total of 1378 observations from 158 candidates were used in the MFRM model, explaining 58.42% of the variance in the data. Fit indices indicated that 1 of the 5 judges inconsistently applied the rating scale. MFRM output also flagged some scores as unexpected based on standardized residual values. This helped identify specific instances where inconsistent observations occurred.

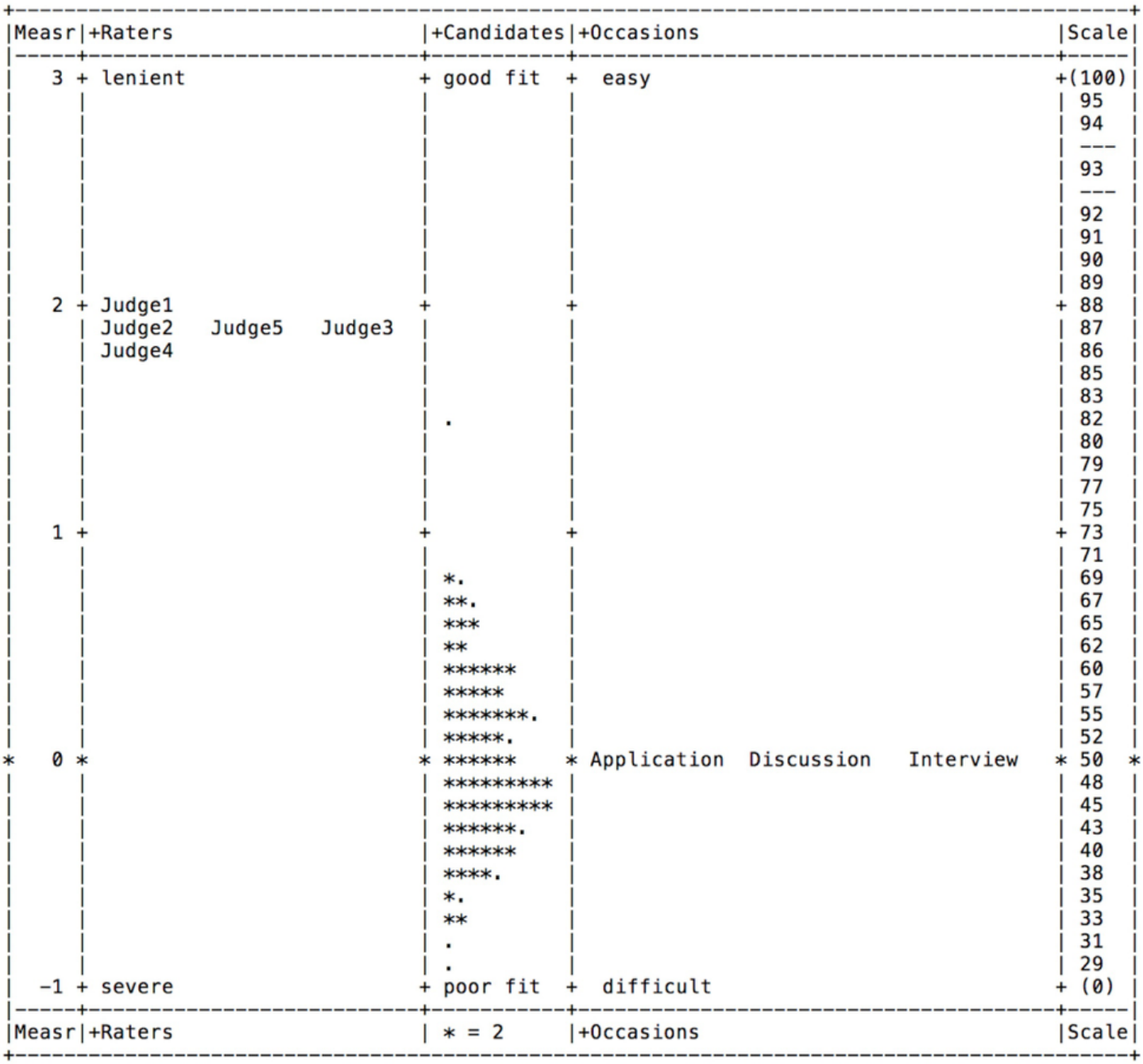
Conclusions: MFRM is a relatively low-cost, efficient way to test the quality of the scores that are used to make a rank list and to investigate noise that represents outlier scores. When these outlier scores are due to biased factors such as particularly stringent or lenient interviewers, they may be unfairly influencing the rank list, and program directors may choose to adjust for them.

Key Words: Residency Application, Residency Interviews, Match List, Psychometrics, Many-facet Rasch measurement.

continued on next page

Figures

Figure 1. *Wright map showing judge leniency, candidate fit, and occasion easiness on a common logit scale (shown in the first column). All three facets are on a positive orientation; higher logit values mean a more lenient judge, a better fit candidate, and easier occasion.*



Figures

Table 1. *Demographics*

Judge	Total Scores Assigned[a]	Observed Average	Fair Average	Leniency, Logits	Standard Error	Infit Mean Squares	Infit *Zstd*[b]
1	470	87.69	88.24	2.02	0.1	1.13	1.8
2	443	86.81	87.36	1.93	0.1	1.04	0.6
3	165	87.61	86.78	1.88	0.2	2.02	6.8
4	171	84.99	86.30	1.84	0.2	1.17	1.5
5	129	86.19	87.47	1.94	0.3	0.98	0.0

[a] These include ratings assigned across 3 time points (occasions).

[b] Z-standardized

Table 2. *Some Unexpected Results Based on Standardized Residuals in Descending Order, the two cases discussed in the text are shaded.*

Candidate ID	Judge	Observed Score	Expected Score	Residual	Standard Residual	Occasion
118	2	65	80.9	−15.9	−4.1	Interview
24	4	75	87.6	−12.6	−3.8	Application
43	3	98	84.8	13.2	3.7	Interview
56	3	88	71.1	16.9	3.7	Interview
17	4	100	89.8	10.2	3.4	Interview
66	1	80	90.2	−10.2	−3.4	Application
34	3	85	93.2	−8.2	−3.2	Interview