

# Rating the Rater: A Technique for Minimizing Leniency Bias in Residency Applications

James R. Seaward, FRCS(Plast.)

Lillian R. Carter, BS

Purushottam Nagarkar, MD

Andrew Y. Zhang, MD

**Background:** Each program in the highly competitive match for a surgical residency needs a way to review applicants effectively. Often this task is undertaken by individual faculty members, reviewing an applicant's file and assigning a score. Despite being asked to rate on a standardized scale, our program found that ratings of the same applicants varied dramatically, with certain faculty consistently scoring higher or lower than others. This is termed leniency bias, or the Hawk-Dove effect, and can affect who is invited to interview depending on which faculty are assigned to review an applicant's file.

**Methods:** A technique to minimize leniency bias was developed and applied to this year's 222 applicants for our plastic surgery residency. The effect of the technique was evaluated by comparing variance between ratings of the same applicants by different faculty before and after our technique.

**Results:** The median variance of ratings of the same applicants reduced from 0.68 before correction to 0.18 after correction, demonstrating better agreement between raters of the applicants' scores after our technique had been applied. This year, applying our technique affected whether or not 16 applicants (36% of interviewees) were invited for interview, including one applicant who matched to our program but who otherwise would not have been offered an interview.

**Conclusions:** We present a simple but effective technique to minimize the leniency bias between raters of residency applicants. Our experience with this technique is presented together with instructions and Excel formulae for other programs to use. (*Plast Reconstr Surg Glob Open* 2023; 11:e4892; doi: [10.1097/GOX.0000000000004892](https://doi.org/10.1097/GOX.0000000000004892); Published online 24 April 2023.)

## INTRODUCTION

Each program in the match for a surgical residency needs a way to review applicants, and often this falls to individual faculty members to assess application files and assign a score to the applicant. This has become more of a challenge for programs in the last few years, as the average number of applications has increased dramatically. In our highly competitive specialty of plastic surgery, the mean number of applications per applicant has increased from 32 in 2018 to 57 in 2022.<sup>1-3</sup> In our program, for the 2022 application cycle, we received 239 applications for five positions in our integrated plastic surgery residency program. While applying initial

screening criteria, including minimum board scores and medical school requirements, this number reduced to 222 applicants, whose applications were reviewed by faculty.

The applications are sent to faculty for review such that each faculty member rates between 15% and 20% of the applications. Any subjective evaluation can be affected by bias, and several types of bias related to residency applications have been described.<sup>4-8</sup> In order to minimize the effect of cognitive bias, each application is rated by multiple faculty and one chief resident.

The applications are scored from one to five, with one being the lowest level of application, and five being the highest. The submitted scores are combined and candidates are ranked in score order, the highest 45 being offered an interview.

Despite being advised to score such that the mean score should be 2.5 and to use the full spread of ratings,

*From the Department of Plastic Surgery, University of Texas Southwestern Medical Center, Dallas, Tex.*

*Received for publication October 21, 2022; accepted February 3, 2023.*

*Copyright © 2023 The Authors. Published by Wolters Kluwer Health, Inc. on behalf of The American Society of Plastic Surgeons. This is an open-access article distributed under the terms of the [Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 \(CCBY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.*

*DOI: [10.1097/GOX.0000000000004892](https://doi.org/10.1097/GOX.0000000000004892)*

Disclosure statements are at the end of this article, following the correspondence information.

Related Digital Media are available in the full-text version of the article on [www.PRSGlobalOpen.com](http://www.PRSGlobalOpen.com).

we found significant variation in both the mean score for each faculty member, and the spread of scores, variance, generated by that faculty member (Fig. 1). This spread is known as the Hawk-Dove effect, or leniency bias. This well-described effect is not readily eliminated by rater training, and raters attempting to correct for their own leniency bias through feedback about their own harshness have been shown to become less consistent in their own ratings than those who continue to behave as they have always done, with lower intra-rater reliability.<sup>9,10</sup>

Due to the large number of applicants to each program, it is common for each applicant to be rated only by a subset of the entire faculty. In our program, for example, we have 20 faculty raters. Our program director rates each application, and each of the others' raters review one-sixth of the applicant pool. This gives an anticipated four ratings of each applicant, the mean of which provides the overall rating of that applicant. In today's competitive environment for surgical residency positions, the leniency or harshness of the faculty to whom each applicant's file is assigned can dramatically affect where they fall in the rank list of applicants, and can influence whether or not they are invited to interview.

Rater bias, as described by Hoyt in his primer article on the subject, refers to disagreements among raters due to either (a) their differential interpretations of the rating scale or (b) their unique (and divergent) perceptions of individual targets.<sup>11</sup> Rater bias can affect the mean of the ratings, the variance of the ratings or the covariance of the ratings with ratings of another attribute. These effects have been traditionally termed "leniency errors" when describing effects that alter the rater's mean, and "halo errors" for effects that alter the covariance. There has been relatively little emphasis on rater variance, despite the fact that this can have substantial effects on rank lists created from raters' scores.<sup>11</sup>

A seemingly simple solution to rater bias would be to have all the raters rate every applicant. Through this complete dataset of ratings of all applicants by all raters, leniency bias is easily identified and eliminated. This, however, is burdensome for our programs and other programs with similar application review structure. The

### Takeaways

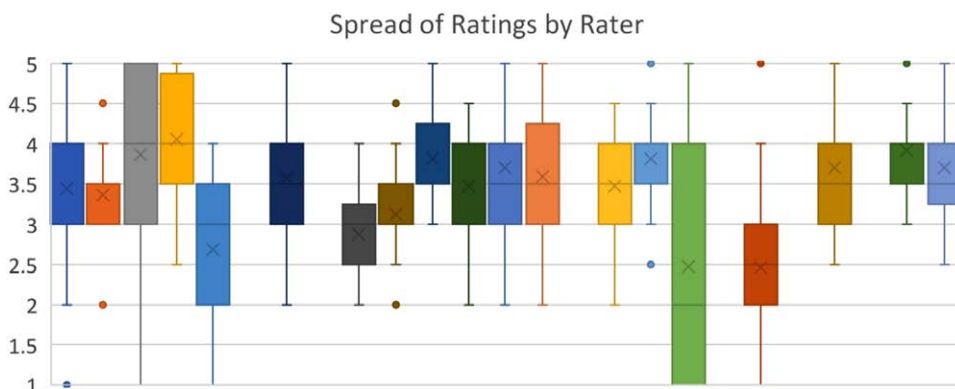
**Question:** Does the application of a normalization technique help address leniency bias in residency application review?

**Findings:** Ratings of the same applicant by different raters had significantly less variance after our normalization process.

**Meaning:** Although bias is not fully eliminated, this technique can help minimize leniency bias in residency application review, and is simple to use.

current assessment process creates an incomplete data set with variable overlap between raters and applicants, leading to difficulties in identifying and correcting rater bias.

Techniques to eliminate rater bias in incomplete data sets have been described.<sup>9,11-13</sup> The majority of this literature is related to standardized examinations, and a review describes the four techniques considered for eliminating this bias from The American College Testing Program, or ACT. The first is through ordinary least squares, a regression-based technique that considers three factors that influence the score given to a candidate: the true ability of the candidate, the leniency bias of the rater, and random error. The technique is simple but assumes that the variance of error across raters is equal. The second technique extends the ordinary least squares model to correct for unequal variance, correcting for each rater's variance through weighting their scoring. The third procedure is an application of the Rasch model, which attempts to model the probability of the applicant's score being correct based on the functions of the applicants' ability and the rater's difficulty. The fourth is through imputed ratings, using established algorithms from multivariate analysis. Using information from individual raters' scores, the incomplete dataset is made complete by estimating or imputing the scores that rater would have provided for all applicants and using multiple regressions to best fill the dataset. The downfall was that the multi-regression style moved all the estimates in the same direction toward



**Fig. 1.** A box and whiskers plot of the spread of ratings given to applicants by each faculty member/ chief resident.

the mean, which may not be the right direction of correction for each rater. This method also requires a lower percentage of missing data in a large data set. Although this model was found to be the optimal solution for the ACT examination, its application for residency applications is, unfortunately, poor due to limitations on the data completeness. The main difference is that in the examination situation, there are many independent scores of the same candidate by the same examiner. These multiple data points for each candidate/examiner pairing can be used to undertake regression analysis or to derive probabilities. For residency application scoring, typically there is a single overall score of the applicant from each rater.

The most widely accepted technique for evaluation and correction for leniency bias in medical settings, the many facet Rasch model, commonly used in objective structured clinical examination and multiple mini-interview environments, again relies on multiple independent scores of the same applicant by the same rater, and is relatively ineffective when only a single rating of each applicant by each rater is provided, as in this situation.<sup>9,10,13,14</sup>

## METHODS

In an attempt to correct for both components of each rater's leniency bias—the mean of the rater's scores and the variance or spread of scores used by each rater, a normalization technique was applied to all ratings. To be able to apply the normalization, the mean and standard deviation were calculated from all available scores for each rater.

Each rating was then normalized using the formula

$$Z = \frac{X - \mu_r}{\sigma_r},$$

where  $Z$  is the normalized value,  $X$  is the raw score,  $\mu_r$  is the rater's mean score, and  $\sigma_r$  is the standard deviation of the rater's scores.

This created a set of scores normalized to a mean of 0 and a standard deviation of 1 for each rater.

Although this normalization would be all that should be required to eliminate the mean and variance leniency biases through this technique, in order to make the scoring more familiar to our faculty, we elected to keep our one-to-five scoring system. As 99.7% of data observed in a normal distribution lies within three standard deviations (SD) of the mean, the values were adjusted such that the mean score was three and a change in score of one represented 1.5 SD above or below the mean: a score of one represented three SD below the mean, two represented 1.5 SD below the mean, three represented the mean, four represented 1.5 SD above the mean and five represented three SD above the mean. This was achieved by multiplying the normalized values by two thirds and then adding 3. Our final normalization formula is as follows:

$$Z = \frac{2}{3} \left( \frac{X - \mu_r}{\sigma_r} \right) + 3.$$

To test whether the scores given by raters for each individual applicant were more consistent after normalization, the variance in rating scores between raters was calculated for each applicant and compared before and after normalization. These values pre- and postcorrection were compared statistically with a Mann-Whitney  $U$  test.

The effects of the correction on the ranking of applicants were evaluated by comparing the rank orders of candidates pre- and postcorrection to determine how individuals moved up and down the lists.

## RESULTS

There were 884 individual application ratings on 222 applicants. Of the 25 raters invited to rate applicants, six raters did not return scores and were excluded from the data set. The program director rated all 222 applications, and the 18 other faculty or chief residents rated 36 or 37 applicants each. The median number of ratings of each applicant by unique raters was four, with a minimum of two and a maximum of five.

Figure 2 demonstrates the score distribution before and after application of our normalization algorithm.

Comparing scores pre- and postcorrection, the sample variance for the set of ratings for each applicant was calculated. The median variance before correction was 0.68, and the median variance after correction was 0.18, resulting in a much smaller difference between ratings of the same applicant after correction. The Mann-Whitney  $U$  test gave a  $U$  value of 7829 with a  $P$  value less than 0.0001. The effect size as defined by  $Z/\sqrt{(n_1+n_2)}$  was large (0.59). The variance distribution pre- and postcorrection is illustrated in Figure 3.

To evaluate the effect of this correction on the applicants, the position of each applicant on the rank list pre- and postcorrection was compared. Table 1 illustrates the change in rank pre-correction versus postcorrection.

As the 45 top-rated applicants were invited to interview, applying this correction on our group of applicants changed whether or not 16 applicants (36% of interviewees) were invited for interview in the 2022 application cycle.

## DISCUSSION

In their publication on best practices in residency application and selection, the Canadian Postgraduate Education Office included the following recommendations: "Selection criteria used for initial filtering, file review, interviews and ranking should be as objective as possible," and "Selection criteria and processes should be fair and transparent for all applicant streams."<sup>15</sup> In an attempt to improve objectivity and fairness, this year our program implemented this normalization process to help reduce rater leniency bias. After normalization, raters tended to agree with one another more on applicant quality, demonstrated by the reduction in variance between ratings of the same applicant following the normalization process.

There are several reports published on how plastic surgery residency programs choose from their applicants in terms of the criteria they find most important, and the

Distribution of Application Scores before and after Normalization

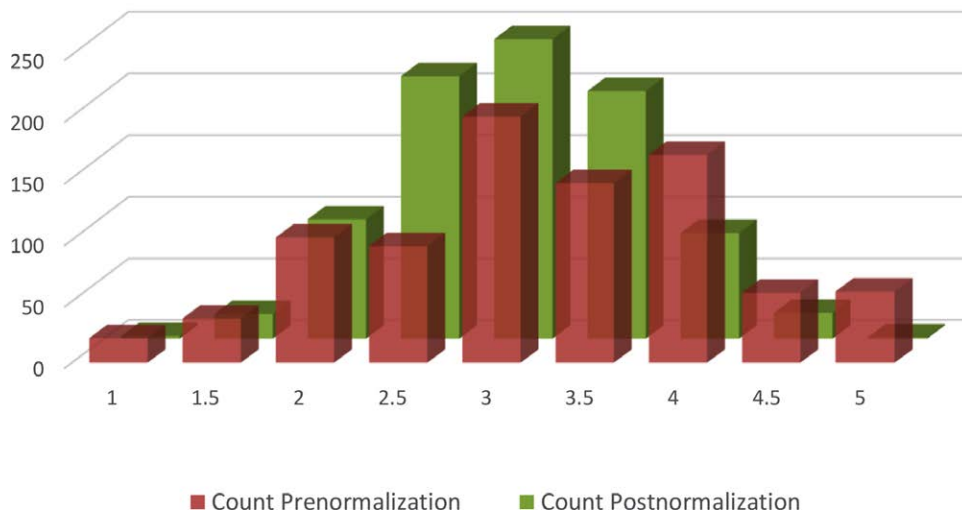


Fig. 2. Bar chart illustrating the total distribution of applicant scores prenormalization (red) and postnormalization (green).

Variance between Ratings of the Same Applicant

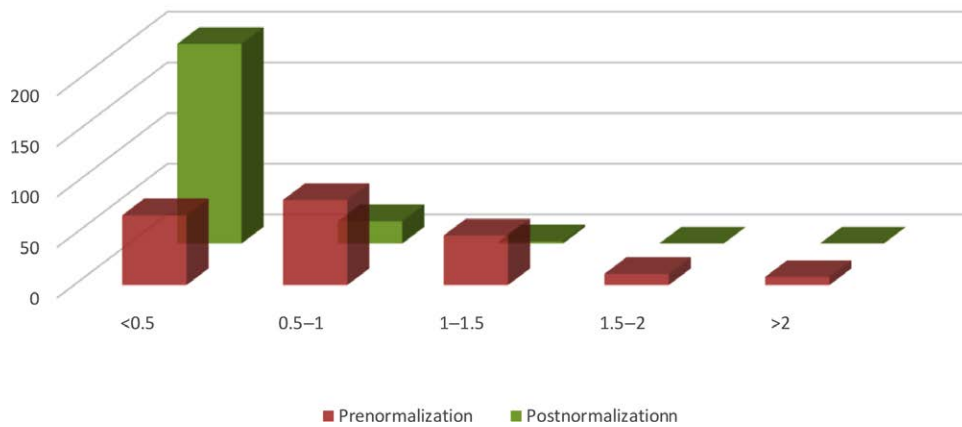


Fig. 3. Bar chart illustrating the variance among scores of the same applicant prenormalization (red) and postnormalization (green).

upcoming elimination of one of those criteria (STEP 1 scores). These reports and articles are typically based on surveys of the program directors of each residency program, and are helpful for future applicants attempting to optimize their application.<sup>16-22</sup> There is little published, however, on how those programs implement their selection criteria, but an initial screening that applicants fulfill minimum requirements followed by individual review by individual raters is common.

If our end point is that ranking should be as objective as possible, we could even consider eliminating the rater: given a set of selection criteria for the residency

and the relative weightings of the importance of those criteria, an algorithm could score applicants in a way that creates a more complete dataset with less rater bias. Of course, although algorithms are promoted as a way to eliminate bias from decision-making, there will still be biases that arise as a result of the selection criteria of the algorithm. One problem with implementing algorithm-based decision-making is that those affected by decisions made by HR algorithms tend to perceive those decisions as less fair than decisions made by humans, even when the outcomes are identical, because they will perceive the algorithmic decision-making process as fundamentally

**Table 1. Illustrating the Extent of Rank Movement and Number of Applicants Affected**

Rank Position Change	Applicant Count
>5	150
>10	113
>15	96
>20	71
>25	53
>30	44
>35	30
>40	19
>45	10
>50	4

reductionistic.<sup>23</sup> These perceptions lead to the assumption that decisions made by algorithms are no different from humans on quantitative input but that they are inferior on qualitative factors, and that overall are based on less complete information and thus are less fair than those made by humans. In addition, several studies have demonstrated that the decision makers prefer to rely more on their own judgment than on linear models and algorithmic decision aids.<sup>23</sup>

Given that some aspects of the application process will be undertaken by human raters for the foreseeable future, how can we change the process to increase fairness? At present, our applicants are sorted alphabetically, and the applicant pool divided into six groups. Each group of applicants is sent to one faculty member with a formal graduate medical education role (chairman, program director of a fellowship program, associate program director of the residency program), three faculty without a formal graduate medical education role, and one chief resident. The program director of the residency program reviews all applicants. Applying the type of correction described in this article is a relatively simple step that can be undertaken retrospectively, but the quality of the leniency bias correction could be improved by implementing a crossed design of which reviewers assess which applicants, and by obtaining more scores per application review: creating individual scores for the applicant's letters of reference, research, STEP scores, honors, extracurriculars, personal statement, etc. This would allow more rigorous techniques of correction, such as the many facet Rasch model to be implemented. We intend to incorporate these changes to our application process for the next cycle.

In the analysis of the effects of our normalization technique, there was a statistically significant difference between the pre- and postnormalization datasets, and there was substantial movement in the ranking list order, with over 100 applicants moving more than 10 positions on the list, and over 70 moving more than 20. Although the top ranked applicants were, in general, still the top ranked applicants pre- and postnormalization, and likewise for those at the bottom of the list, there was more movement in the list order below the top 30. We invited 45 applicants for interview and, in the top 45, sixteen applicants (36% of interviewees) were either invited for interview, or not invited for interview as a result of our normalization

process, including one applicant who ended up matching into our program.

Although this formula has demonstrated significant reduction in variance in the preinterview phase, there are still areas of bias and variability along the path from the first day of medical school to the match. In addition, there are curriculum changes that have impacts that have yet to be uncovered. For example, this year (2022) is the first year that Step 1 is graded with a pass or fail score. This is a significant change to a score that has historically been a major determining factor for residency qualifications. This will result in Step 2 CK scores taking on greater weight. There will also be a greater emphasis on applicant familiarity, away rotations, and letters of recommendation, increasing the in-group bias and geographical bias already described in residency applications.<sup>4,6,7,16</sup> This further forces students to make career specialty decisions even earlier, and great applicants may be overlooked if they develop their passions later than most.

This technique is not without assumptions and limitations. Probably the largest is the assumption that the groups of 36 to 37 applicants reviewed by each faculty member are equivalent and so correcting each group to the same mean and standard deviation is valid. Also, variance may not be the optimal outcome measure to evaluate the quality of this process when the process itself is designed to make changes to the rater's variance. In defense of this outcome measure, although the mean variance for each rater dropped from 0.62 to 0.44 (29%) with the normalization process the mean variance by applicant dropped from 0.76 to 0.23 (74%). In future years, we hope to improve on this outcome measure by using the individual elements of the many facet Rasch model.

In conclusion, we propose a simple technique for normalization of the ratings of residency applicants, which in our applicant group this year demonstrated reduced variance between ratings of the same applicant by different faculty. In Supplemental Digital Content 1, we include detailed instructions for interested programs to be able to implement this technique. (See appendix, Supplemental Digital Content 1, which displays the instructions for programs to implement this technique in Excel. <http://links.lww.com/PRSGO/C467>.)

James R. Seaward, MD

1801 Inwood Rd

Dallas, TX, 75390

E-mail: [james.seaward@utsouthwestern.edu](mailto:james.seaward@utsouthwestern.edu)

#### DISCLOSURE

*The authors have no financial interest to declare in relation to the content of this article.*

#### REFERENCES

1. ERAS Statistics. Association of American Medical Colleges. Available at <https://www.aamc.org/data-reports/interactive-data/eras-statistics-data>
2. Asserson DB, Sarac BA, Janis JE. A 5-year analysis of the integrated plastic surgery residency match: the most competitive specialty? *J Surg Res.* 2022;277:303–309.

3. Sarac BA, Janis JE. Matching into plastic surgery: insights into the data. *Plast Reconstr Surg Glob Open*. 2022;10:e4323.
4. Bass A, Wu C, Schaefer JP, et al. In-group bias in residency selection. *Med Teach*. 2013;35:747–751.
5. Fuchs JW, Youmans QR. Mitigating bias in the era of virtual residency and fellowship interviews. *J Grad Med Edu*. 2020;12:674–677.
6. Nagarkar PA, Janis JE. Unintended bias and unintended consequences: geographic bias in the plastic surgery residency match. *Plast Reconstr Surg Glob Open*. 2022;10:e4063.
7. Ryan T. Addressing bias and lack of objectivity in the Canadian resident matching process. *CMAJ*. 2018;190:E1211–E1212.
8. Griffin BN, Wilson IG. Interviewer bias in medical student selection. *Med J Aust*. 2010;193:343–346.
9. 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 modeling. *BMC Med Educ*. 2006;6:42.
10. Roberts C, Rothnie I, Zoanetti N, et al. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Med Educ*. 2010;44:690–698.
11. Hoyt WT. Rater bias in psychological research: when is it a problem and what can we do about it? *Psychol Methods*. 2000;5:64–86.
12. Raymond MR, Houston WM. Detecting and correcting for rater effects in performance assessment. Vol. 90-14. December 1990. Available at [https://www.act.org/content/dam/act/unsecured/documents/ACT\\_RR90-14.pdf](https://www.act.org/content/dam/act/unsecured/documents/ACT_RR90-14.pdf)
13. Kiraly L, Dewey E, Brasel K. Hawks and doves: adjusting for bias in residency interview scoring. *J Surg Educ*. 2020;77:e132–e137.
14. Yeates P, Sebok-Syer SS. Hawks, doves and rasch decisions: understanding the influence of different cycles of an OSCE on students’ scores using many facet rasch modeling. *Med Teach*. 2017;39:92–99.
15. Bandiera G, Abrahams C, Ruetalo M, et al. Identifying and promoting best practices in residency application and selection in a complex academic health network. *Acad Med*. 2015;90:1594–1601.
16. Asaad M, Drolet BC, Janis JE, et al. Applicant familiarity becomes most important evaluation factor in USMLE step i conversion to pass/fail: a survey of plastic surgery program directors. *J Surg Edu*. 2021;78:1406–1412.
17. Janis JE, Hatef DA. Resident selection protocols in plastic surgery: a national survey of plastic surgery program directors. *Plast Reconstr Surg*. 2008;122:1929–1939.
18. Nguyen AT, Janis JE. Resident selection protocols in plastic surgery: a national survey of plastic surgery independent program directors. *Plast Reconstr Surg*. 2012;130:459–469.
19. Krauss EM, Bezuhly M, Williams JG. Selecting the best and brightest: a comparison of residency match processes in the United States and Canada. *Plast Surg (Oakv)*. 2015;23:225–30.
20. Nagarkar P, Pulikkottil B, Patel A, et al. So you want to become a plastic surgeon? What you need to do and know to get into a plastic surgery residency. *Plast Reconstr Surg*. 2013;131:419–422.
21. Lin LO, Makhoul AT, Hackenberger PN, et al. Implications of pass/fail step 1 scoring: plastic surgery program director and applicant perspective. *Plast Reconstr Surg Glob Open*. 2020;8:e3266.
22. Raborn LN, Janis JE. Current views on the new united states medical licensing examination step 1 pass/fail format: a review of the literature. *J Surg Res*. 2022;274:31–45.
23. Newman DT, Fast NJ, Harmon DJ. When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organ Behav Human Decis Process*. 2020;160:149–167.