9-8-2023

Hawks and Doves: Perceptions and Reality of Faculty Evaluations

Jillian Zavodnick
Thomas Jefferson University

Jonathan Doroshow

Sarah Rosenberg
Thomas Jefferson University

Joshua Banks

Benjamin E. Leiby

See next page for additional authors

Hawks and Doves: Perceptions and Reality of Faculty Evaluations

Journal of Medical Education and Curricular Development
Volume 10: 1–7

DOI: 10.1177/23821205231197079

Jillian Zavodnick[1], Jonathan Doroshow[2], Sarah Rosenberg[1], Joshua Banks[3], Benjamin E Leiby[3] and Nina Mingioni[1]

[1]Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, USA. [2]Department of Medicine, Lankenau Medical Center, Wynnewood, USA. [3]Department of Pharmacology and Experimental Therapeutics, Division of Biostatistics, Thomas Jefferson University, Philadelphia, USA.

ABSTRACT

OBJECTIVES: Internal medicine clerkship grades are important for residency selection, but inconsistencies between evaluator ratings threaten their ability to accurately represent student performance and perceived fairness. Clerkship grading committees are recommended as best practice, but the mechanisms by which they promote accuracy and fairness are not certain. The ability of a committee to reliably assess and account for grading stringency of individual evaluators has not been previously studied.

METHODS: This is a retrospective analysis of evaluations completed by faculty considered to be stringent, lenient, or neutral graders by members of a grading committee of a single medical college. Faculty evaluations were assessed for differences in ratings on individual skills and recommendations for final grade between perceived stringency categories. Logistic regression was used to determine if actual assigned ratings varied based on perceived faculty's grading stringency category.

RESULTS: "Easy graders" consistently had the highest probability of awarding an above-average rating, and "hard graders" consistently had the lowest probability of awarding an above-average rating, though this finding only reached statistical significance only for 2 of 8 questions on the evaluation form ($P = .033$ and $P = .001$). Odds ratios of assigning a higher final suggested grade followed the expected pattern (higher for "easy" and "neutral" compared to "hard," higher for "easy" compared to "neutral") but did not reach statistical significance.

CONCLUSIONS: Perceived differences in faculty grading stringency have basis in reality for clerkship evaluation elements. However, final grades recommended by faculty perceived as "stringent" or "lenient" did not differ. Perceptions of "hawks" and "doves" are not just lore but may not have implications for students' final grades. Continued research to describe the "hawk and dove effect" will be crucial to enable assessment of local grading variation and empower local educational leadership to correct, but not overcorrect, for this effect to maintain fairness in student evaluations.

KEYWORDS: Assessment, evaluation, faculty, clerkship, grade, grading committee, internal medicine

RECEIVED: July 27, 2023. **ACCEPTED:** August 8, 2023

TYPE: Original Research Article

FUNDING: This work did not receive external support, and was funded internally by Sidney Kimmel Medical College's Department of Medicine. Publication made possible in part by support from the Thomas Jefferson University Open Access Fund.

DECLARATION OF CONFLICTING INTERESTS: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Nina Mingioni, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, USA. Email: nina.mingioni@jefferson.edu

Introduction

Internal medicine (IM) clerkship grades, and their derivatives such as class rank and honorary society memberships, are an important factor in the residency selection process.[1,2] With the transition of the United States Medical Licensing Exam (USMLE) Step 1 to pass/fail scoring, the importance of clerkship grades will likely increase further.[1,2] There is great variability in grading processes among medical schools, such as different grading tiers, percentage of high grades, processes for assigning grades, and components contributing to the grades.[3–8] However, 90% of institutions use some form of clinical performance assessment in the calculation of a clerkship grade, comprising, on average, 52.8% of that final grade.[9] Though some amount of subjectivity is inherent and perhaps even desirable for the assessment of clinical performance,[10] there are concerns about the reliability of evaluator data,[11–13] as well as differences between evaluators assessing the same performance.[14] Additionally, narrative comments use different language to describe students based on gender and underrepresented-in-medicine status,[15–17] with students from racial and ethnic groups underrepresented in medicine systematically receiving lower clerkship grades.[15,18,19] It is unsurprising that students do not perceive clerkship grades as fair and accurate.[20]

Clerkship grading committee review of individual evaluator data for determination of final grade is recommended as a method to improve the consistency of high-stakes decision making, increase detection of poor performance, and minimize the effect of individual evaluator bias on final grade.[21–23] The proposed mechanism of these advantages is a shared mental model of performance expectations, reliable interpretation of available data (clerkship evaluations), and an understanding of shared accountability.[24] An understanding of individual evaluator characteristics could improve interpretation of data from that evaluator,[25,26] but the ability of a committee to assess evaluator stringency or leniency has not been previously

studied. Knowing whether any given evaluator has a tendency to be a stringent grader (a "hawk") or a lenient grader (a "dove") provides a context for the grading committee members as to how the evaluator Likert ratings or comments can be interpreted as compared to others.

We hypothesized that an experienced grading committee can reliably detect patterns of stringency or leniency when exposed to evaluator data over time. We aimed to evaluate a statistical reality of our clerkship grading committee's assessment of evaluators as "hawks" or "doves."

Methods

Setting and study population

This work was conducted at a large medical school (average class size 276) with several affiliated clinical sites. The IM clerkship is a mandatory 8-week course completed during the third year of medical school at the time of the study. Four weeks of clerkship are completed at the main university hospital, and four weeks are completed at an affiliated site—an academic community medical center.

This study analyzed student evaluations for the 2017-18 and 2018-19 academic years at the Sidney Kimmel Medical College (SKMC). At that time, the clerkship grades were comprised of a score based on students' clinical performance and other factors including the NBME (National Board of Medical Examiners) subject examination and special projects. For each student, faculty and housestaff who worked with the student in the clinical setting during the clerkship completed a standard Clerkship Evaluation Form. This form is used for every clerkship at the medical college and can be found in supplemental materials (Clerkship Evaluation Form). It consists of two narrative fields, soliciting summative and formative feedback, as well as 8 questions based on competencies and SKMC's medical education program objectives, each rated by evaluators on a 3-point Likert scale (below expected, expected, and above expected). The questions are displayed in Table 1. At the end, each evaluator is asked to recommend a final grade for the student based on their clinical performance. For the academic years included in the analysis, grade choices were Honors, Excellent, Good-Plus, Good, Good-minus, Marginal, and Failure. For each student, evaluations are collected electronically then combined into a single digital composite. A web-based evaluation management software called *New Innovations* (https://www.new-innov.com) is utilized for this process.

The IM grading committee is comprised of the clerkship director, all clinical affiliate clerkship site directors, as well as other members of the IM Undergraduate Medical Education (UME) leadership. The grading committee meets at the end of each block to review all evaluations students receive. Members review the summative narratives, as well as the Likert items. Through a discussion, a consensus is achieved,

Table 1. Quantitative fields of the Clerkship Evaluation Form completed by evaluators for each student.

Question number	Question text	Rating options
1	Ability to establish humanistic rapport with patient. Ability to gather essential and accurate information about patients and their conditions through history-taking and physical examination.	Below expected Expected Above expected
2	Demonstrates appropriate knowledge base and understanding of diseases. Uses evidence-based medicine. Applies knowledge in clinical situations and constructs a differential diagnosis. Formulates a treatment plan.	
3	Able to identify own strengths and areas for improvement. Able to accept feedback, and incorporate it into daily practice of medicine to improve own performance.	
4	Able to communicate with team about clinical, administrative, and personal tasks. Ability to report data in both oral and written form in clear, succinct, and organized manner. Able to maintain a clear, legible, and appropriate medical record. Able to engage patients in education.	
5	Able to demonstrate compassion, integrity, and respect for others. Demonstrates sensitivity and responsiveness to a diverse patient population. Demonstrates integrity and commitment to ethical principles. Respects patient confidentiality.	
6	Able to effectively utilize available resources. Advocates for patient safety. Aware of concepts of cost, quality, and patient safety.	
7	Works with other health professionals and staff to establish and maintain a climate of mutual respect.	
8	Demonstrates personal accountability. Manages competing needs of personal and professional responsibility. Demonstrates trustworthiness to one's colleagues regarding the care of patients.	
	Final grade	Fail Marginal Good – Good Good + Excellent Honors

and each student receives a numerical score commensurate with their clinical performance. This score is then used in calculating the final grade for each student.

During the group discussions, the narrative comments are weighed most heavily, although Likert items and each evaluator's final suggested grade are considered. The narrative comments were not analyzed in this study. Quantitative ratings were selected for this analysis due to the ease of quantitative analytics available to the study team. Since the pool of evaluators is finite, each evaluator is known to the members of the grading committee. Additionally, with each faculty member evaluating multiple students over a course of several academic years, patterns in how they fill out evaluations, as well as the language they use in their narratives, are well-known to the grading committee members. In reviewing and interpreting evaluators' comments and Likert scale ratings, the grading committee considers the level of experience of the evaluator as well as the grading committee members' global assessment of the evaluator's stringency, assessment acumen, and dedication to the evaluation process – an assessment sometimes informally described among the committee as an evaluator being a "hawk," a "dove," or "spot on." Four of the authors (JZ, SR, JD, and NM) have been members of the Grading Committee for a minimum of 3 years (JZ and SR), and a maximum of 12 years (NM).

Data collection

We extracted evaluation data from the evaluation management software for all IM clerkships completed during the 2017-18 and 2018-19 academic years. The data set consisted of evaluator name, clinical site, and quantitative data available on each evaluation they completed—answers to all Likert-rated questions and evaluator's suggested final grade. Student names, the final grades assigned by the grading committee, and narrative comments were not included in this analysis.

From this data, a convenience sample was created, including only evaluations completed by the faculty at the Thomas Jefferson University Hospital (TJUH) and one of its academic affiliates, Lankenau Medical Center (LMC), due to the multiple authors' familiarity with the grading patterns of that faculty. Housestaff evaluations of students were excluded. From this sample, a list of faculty names who completed at least one student evaluation during the study period was generated. Each faculty member on the list was rated based on the grading stringency impression as a "hard," "neutral," or "easy" grader. One author (JD), who is on staff at LMC, assigned grading stringency categories based on his impressions of faculty at his site; other authors (JZ, SR, and NM) concurred with his assessments. Twenty-three affiliate faculty members completed an evaluation during the study period; all were included in the analysis.

Three authors (JZ, SR, and NM) are on staff at TJUH, where 66 faculty members completed at least one clerkship evaluation during the study period. Each TJUH author assigned ratings to all faculty members for whom they had an impression of grading stringency. Faculty were excluded if zero or only one author had an impression ($n = 18$ and $n = 17$, respectively). One faculty member was excluded due to inability to separate evaluations that individual completed as a senior resident in the first year of the study period form those completed as a faculty in the second year of the study period. Impressions were adopted if three authors had a consensus on the stringency impression, or if two had the same impression and the third had no impression. If one author disagreed with the other two, or two authors had different impressions, the impression was discussed; if consensus could not be achieved, the faculty member was excluded ($n = 3$). For one faculty member, the authors had widely different impressions (one judged the faculty member a "hard" grader, another an "easy" grader) and this faculty member was excluded without discussion. After exclusions, evaluations from 27 TJUH faculty were analyzed.

Statistical analysis

The final sample included in the analysis consisted of 877 student evaluations by 52 faculty members. For each evaluation question, binary logistic regression was used to determine if the odds of having an above-average assigned rating varied based on perceived faculty stringency category. Ordinal logistic regression was used to model the odds of suggesting a certain final clerkship grade. For all logistic regression models, generalized estimating equations were used to account for correlation of grading within each individual faculty grader.

Raw data used for this analysis is available up on request. The Thomas Jefferson University Institutional Review Board evaluated this study and determined it to be exempt from review, waiving the requirement to obtain informed consent.

Results

Table 2 shows the probability of faculty in each stringency category assigning an above-average rating for each of the eight Likert scale questions. Although "easy graders" consistently had the highest probability of awarding an above-average rating, and "hard graders" consistently had the lowest probability of awarding an above-average rating, this finding only reached statistical significance for questions 2 (rating knowledge base, use of evidence-based medicine, application of knowledge in clinical situations, differential diagnoses, and treatment plans; $P = .033$) and 3 (rating the ability identify own strengths and area of improvement, acceptance of feedback and performance improvement; $P = .001$).

"Easy graders" had a significantly higher odds ratio (OR) of giving an above-average rating as compared to "hard

Table 2. Probabilities and odds ratios of giving an above-average rating for questions 1–8 based on stringency impression.

Question #	Probability of an above average rating			OR (95% CI) of giving an above-average rating			*P*-value
	Easy	Neutral	Hard	Easy versus Neutral	Hard versus Neutral	Easy versus Hard	
1	0.86	0.84	0.84	1.13 (0.98-1.31)	0.98 (0.85-1.13)	1.16 (0.99-1.37)	.142
2	0.64	0.48	0.44	2.00 (1.06-3.79)	0.86 (0.51-1.45)	2.34 (1.22-4.48)	.033
3	0.78	0.61	0.55	2.20 (1.08-4.50)	0.77 (0.40-1.47)	2.87 (1.44-5.71)	.001
4	0.76	0.64	0.57	1.81 (0.86-3.82)	0.77 (0.42-1.39)	2.36 (1.12-4.97)	.078
5	0.78	0.73	0.68	1.36 (0.57-3.25)	0.81 (0.38-1.97)	1.67 (0.68- 4.09)	.534
6	0.52	0.36	0.33	1.91 (0.79-4.67)	0.86 (0.39-1.92)	2.23 (0.92-5.39)	.1845
7	0.77	0.65	0.68	1.79 (0.79-4.08)	1.12 (0.48-2.58)	1.60 (0.65-3.93)	.359.
8	0.83	0.67	0.65	2.39 (1.026-5.55)	0.942 (0.031-1.827)	2.53 (1.03-6.22)	.077

Table 3. Odds ratios for assigning a higher final rating compared to another stringency category.

	OR (95% CI)			*P*-value
	Easy versus Neutral	Hard versus Neutral	Easy versus Hard	
Final rating	1.54 (0.69-3.43)	0.753 (0.34-1.67)	2.04 (0.81-5.14)	.3069

graders" for questions 2 (OR 2.34, 95% CI 1.22-4.48), 3 (OR 2.87, 95% CI 1.44-5.71), 4 (OR 2.36, 95% CI 1.12-4.97), and 8 (OR 2.53, 95% CI 1.03-6.22). Comparing "easy" to "neutral" graders, the OR of giving an above-average rating was significant for questions 2 (OR 2.00, 95% CI 1.06-3.79), 3 (OR 2.20, 95% CI 1.08-4.50), and 8 (OR 2.39, 95% CI 1.026-5.55). Odds ratios for other questions and other stringency comparisons were not significantly different.

Odds ratio for assigning a higher suggested final grade for each stringency category compared to the others is shown in Table 3. None of these ORs were significant. Distribution of final ratings for each stringency category of evaluator is shown in Table 4 and is displayed in Figure 1. For all categories of grading stringency, the most common suggested final grade was "excellent," followed by "honors." "Hard graders" had a smaller portion of "Honors" among their final suggested grades as compared to "neutral" graders, who in turn had a smaller portion of "honors" among their final suggested grades as compared to "easy" graders. This pattern was also true for a final suggested grade of "excellent." "Good-plus" was a larger proportion of suggested grades within the category of "neutral" graders compared to "hard," and of "hard" compared to "easy." "Good" and "good-minus" were a larger portion of grades for "hard" graders than "neutral," and of "neutral" than "easy." "Good-minus" was rarely assigned,

Table 4. Distribution of final ratings for each stringency category of evaluator.

Grade	% (n)		
	Easy	Neutral	Hard
Honors	36.59% (105)	28.78% (80)	25.11% (55)
Excellent	45.30% (130)	43.53% (121)	41.55% (91)
Good +	14.29% (41)	19.42% (54)	18.26% (40)
Good	3.48% (10)	5.40% (15)	11.42% (25)
Good –	0.35% (1)	2.16% (6)	2.28% (5)
Marginal	0.00% (0)	0.72% (2)	1.37% (3)

and only once by an easy grader; "marginal" was rarely assigned, and never by an easy grader.

Discussion

This analysis demonstrates statistically significant differences in ratings assigned by evaluators perceived by experienced grading committee members to be particularly stringent or lenient in their evaluations. Evaluation items related to three domains of competency[27] demonstrated a statistically significant difference between perceived hard, neutral, and easy graders. A clear difference existed for questions 2 and 3, which assessed knowledge for practice, patient care, and practice-based learning and improvement. Skills mapping to more subjectively assessed domains, like interpersonal and communication skills and professionalism, did not show any significant difference within our study group. There was also no difference in the final suggested grades the evaluators recommended. Despite the lack of statistical significance in many of the evaluation ratings, the expected pattern of outcomes was nearly universal, raising the possibility of type II error.

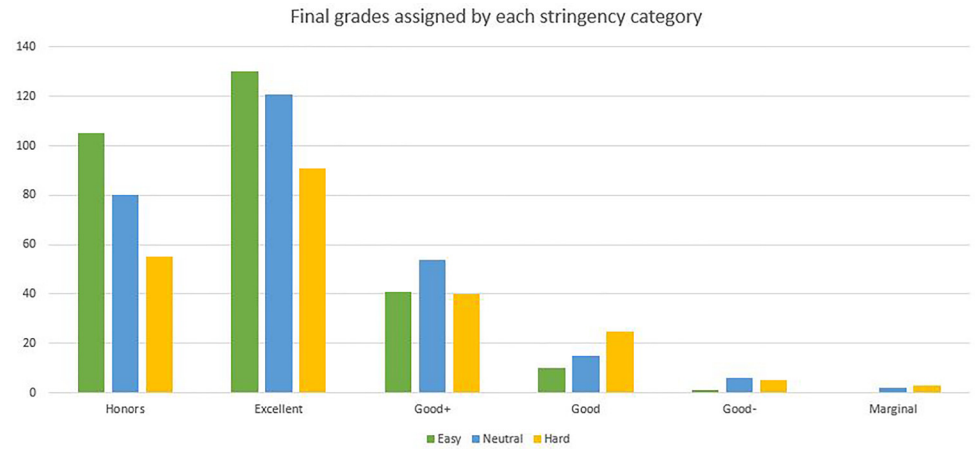

Figure 1. Distribution of final suggested grades assigned by easy, neutral, and hard graders.

A general perception exists among medical students that the clerkship grading process is unfair.[20] Identifiable differences in the grading patterns of faculty that complete evaluations may contribute to such opinions, and this perception is borne out by imperfect inter-rater correlation when assessing the same performance.[14] Interestingly, prior studies demonstrated that efforts to train the evaluators in order to improve inter-rater reliability are generally not successful.[28,29] Seasoned clerkship directors often have insight into the variability of expectations of medical students among their faculty. These perceptions of faculty are drawn from commentary on student evaluations, as well as various degrees of personal knowledge of the faculty members. Despite similar final grade recommendations across stringency categories, the existence of differences in assessing discrete skills demonstrates that this grading committee's perception of stringency and leniency has some basis in reality. However, the small overall differences, and the lack of detectable difference in final suggested grade, should provide some reassurance to students that being assigned to a hard-grading "hawk" faculty is not a guarantee of a lower grade.

This analysis has limitations. First, narrative comments were not included in the dataset. Many medical schools rely heavily on commentary provided by faculty to determine a final grade for students. A qualitative analysis, examining the differences in narrative comments by perceived grading stringency or leniency, is an important next step to support the existence of perceivable "hawks" (hard graders) and "doves" (easy graders). Perhaps such an investigation could focus on comments reflecting achievement in knowledge for practice, patient care, and practice-based learning and improvement, the domains found in our study to have detectable differences between evaluator categories. This could be particularly critical for schools with a pass-fail clerkship grading system, where evaluator comments may be the only data in the Medical School Performance Evaluation or departmental letter of support that can discriminate between the levels of student achievement; in that case, assignment to a "hawk" in the clerkship could have serious consequences for residency selection.

Another limitation is the inability to connect evaluations with the final grade assigned to the student. Though our findings suggest that some level of clerkship director and grading committee compensation for faculty tendencies is appropriate, we cannot determine if that level of compensation is appropriate. A finding that students are equally likely to obtain a given final grade regardless of the composition of "hawks" and "doves" among their evaluators would demonstrate appropriate grading committee adjustment for these categories when assigning final grade. However, over- or under-compensation is also possible. Though this study demonstrates a reality underlying the categories "hawk" and "dove," it cannot determine whether this grading committee reacts appropriately to these perceptions.

Additionally, a power analysis was not performed; instead, all available evaluations that met eligibility criteria were included to obtain the largest possible sample size. This limits the reliability of our negative findings. The Clerkship Evaluation Form used to obtain data is not a validated instrument. Although it was not formally pilot tested, it had been in use at our medical school for over a year prior to data collection.

There are other variables that may have impacted our study. Medical students have a wide range of skill levels. We were not able to control for the skill level of the students assigned to a specific evaluator. It is possible some faculty worked with a disproportionate number of lower-performing students compared to their peers, and their "hawk" reputation rests not on the unusually high expectations but rather on unusually weaker students. The affiliate site faculty work with learners from several different medical schools, which may further impact variability in expectations and perception of students from the one medical school studied. Different numbers of contact hours between an individual student and evaluator might have influenced assessments, but this variable was not available.

Lastly, we conducted this assessment within a single grading committee. More research is needed to determine if the ability to detect evaluator stringency is a universal or even a common feature of grading committees, and whether features of a grading committee (eg, meeting structure, membership composition) or its individual members (eg, years of experience in education or on the committee, time dedicated to education) contribute to accurate or inaccurate assessment of evaluator stringency.

Despite these limitations, our findings support a factual basis for perception of evaluators as "hawks" or "doves." Clerkship directors and those who serve on a grading committee may wish to compare their impressions of the individual evaluators, develop a shared mental model of the "hawks" and "doves," and perform some form of analysis, meaningful to their grading procedures, to ensure that any compensation for stringency undertaken by their committee is appropriate, such as the variance analyses of Generalizability theory[30] or the method described by Murphy et al.[31]

Conclusions

Perceived differences in faculty grading stringency have basis in reality for individual clerkship evaluation elements. However, final grades recommended by faculty perceived as stringent or lenient did not differ. Perceptions of "hawks" and "doves" are not just lore but may not have implications for the final grades. Continued research to describe the "hawk and dove effect" will be crucial to enable assessment of local grading variation and empower local educational leadership to correct, but not overcorrect, for this effect to maintain fairness in student grading. This will become increasingly crucial as factors previously helpful for residency selection, such as the USMLE Step 1, and even clerkship grades, become increasingly pass/fail.

Acknowledgement

The authors would like to acknowledge Amanda White for her assistance with data collection.

ORCID iD

Nina Mingioni https://orcid.org/0000-0001-7350-8604

Supplemental material

Supplemental material for this article is available online.

REFERENCES

1. National Resident Matching Program, Data Release and Research Committee: Results of the 2018 NRMP Program Director Survey. National Resident Matching Program, Washington, DC.2018. https://www.nrmp.org/wp-content/uploads/2018/07/NRMP-2018-Program-Director-Survey-for-WWW.pdf. Accessed February 21, 2021.
2. Hartman ND, Lefebvre CW, Manthey DE. A narrative review of the evidence supporting factors used by residency program directors to select applicants for interviews. *J Grad Med Educ*. 2019;11(3):268-273. doi:10.4300/JGME-D-18-00979.3
3. Takayama H, Grinsell R, Brock D, Foy H, Pellegrini C, Horvath K. Is it appropriate to use core clerkship grades in the selection of residents? *Curr Surg*. 2006;63(6):391-396. doi:10.1016/j.cursur.2006.06.012
4. Alexander EK, Osman NY, Walling JL, Mitchell VG. Variation and imprecision of clerkship grading in U.S. medical schools. *Acad Med*. 2012;87(8):1070-1076. doi:10.1097/ACM.0b013e31825d0a2a
5. Fazio SB, Torre DM, DeFer TM. Grading practices and distributions across internal medicine clerkships. *Teach Learn Med*. 2016;28(3):286-292. doi:10.1080/10401334.2016.1164605
6. Grading Systems Use by US Medical Schools | AAMC. https://www.aamc.org/data-reports/curriculum-reports/interactive-data/grading-systems-use-us-medical-schools. Accessed February 22, 2021.
7. Association of American Medical Colleges. Grading Systems Used by US Medical Schools. AAMC Curriculum Inventory, 2015-2020. https://www.aamc.org/data-reports/curriculum-reports/interactive-data/grading-systems-use-us-medical-schools. Accessed November 24, 2021.
8. Hemmer PA, Papp KK, Mechaber AJ, Durning SJ. Evaluation, grading, and use of the RIME vocabulary on internal medicine clerkships: results of a national survey and comparison to other clinical clerkships. *Teach Learn Med*. 2008;20(2):118-126. doi:10.1080/10401330801991287
9. Hernandez CA, Daroowalla F, LaRochelle JS, et al. Determining grades in the internal medicine clerkship: results of a national survey of clerkship directors. *Acad Med*. 2021;96(2):249-255. doi:10.1097/ACM.0000000000003815
10. Ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2019;94(3):333-337. doi:10.1097/ACM.0000000000002495
11. Hauer KE, Lucey CR. Core clerkship grading: the illusion of objectivity. *Acad Med*. 2019;94(4):469-472. doi:10.1097/ACM.0000000000002413
12. Lye PS, Biernat KA, Bragg DS, Simpson DE. A pleasure to work with—an analysis of written comments on student evaluations. *Ambul Pediatr*. 2001;1(3):128-131.
13. Jackson JL, Kay C, Jackson WC, Frank M. The quality of written feedback by attendings of internal medicine residents. *J Gen Intern Med*. 2015;30(7):973-978. doi:10.1007/s11606-015-3237-2
14. 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 Health Sci Educ Theory Pract*. 2013;18(3):325-341. doi:10.1007/s10459-012-9372-1
15. Rojek AE, Khanna R, Yim JWL, et al. Differences in narrative language in evaluations of medical students by gender and under-represented minority status. *J Gen Intern Med*. 2019;34(5):684-691. doi:10.1007/s11606-019-04889-9
16. Axelson RD, Solow CM, Ferguson KJ, Cohen MB. Assessing implicit gender bias in medical student performance evaluations. *Eval Health Prof*. 2010;33(3):365-385. doi:10.1177/0163278710375097
17. Gorth DJ, Magee RG, Rosenberg SE, Mingioni N. Gender disparity in evaluation of internal medicine clerkship performance. *JAMA Netw Open*. 2021;4(7):e2115661. doi:10.1001/jamanetworkopen.2021.15661
18. Low D, Pollack SW, Liao ZC, et al. Racial/ethnic disparities in clinical grading in medical school. *Teach Learn Med*. 2019;31(5):487-496. doi:10.1080/10401334.2019.1597724
19. Teherani A, Hauer KE, Fernandez A, King TE, Lucey C. How small differences in assessed clinical performance amplify to large differences in grades and awards: a cascade with serious consequences for students underrepresented in medicine. *Acad Med*. 2018;93(9):1286-1292. doi:10.1097/ACM.0000000000002323
20. Bullock JL, Lai CJ, Lockspeiser T, et al. In pursuit of honors: a multi-institutional study of students' perceptions of clerkship evaluation and grading. *Acad Med*. 2019;94(11S). Association of American Medical Colleges Learn Serve Lead: Proceedings of the 58th Annual Research in Medical Education Sessions:S48-S56. doi:10.1097/ACM.0000000000002905
21. Frank AK, O'Sullivan P, Mills LM, Muller-Juge V, Hauer KE. Clerkship grading committees: the impact of group decision-making for clerkship grading. *J Gen Intern Med*. 2019;34(5):669-676. doi:10.1007/s11606-019-04879-x
22. Onumah CM, Lai CJ, Levine D, Ismail N, Pincavage AT, Osman NY. Aiming for equity in clerkship grading: recommendations for reducing the effects of structural and individual bias. *Am J Med*. 2021;134(9):1175-1183.e4. doi:10.1016/j.amjmed.2021.06.001
23. Hauer KE, Cate OT, Boscardin CK, et al. Ensuring resident competence: a narrative review of the literature on group decision making to inform the work of clinical competency committees. *J Grad Med Educ*. 2016;8(2):156-164. doi:10.4300/JGME-D-15-00144.1
24. Edgar L, Jones MD, Harsy B, Passiment M, Hauer KE. Better decision-making: shared mental models and the clinical competency committee. *J Grad Med Educ*. 2021;13(2 Suppl):51-58. doi:10.4300/JGME-D-20-00850.1
25. Ekpenyong A, Baker E, Harris I, et al. How do clinical competency committees use different sources of data to assess residents' performance on the internal medicine milestones? A mixed methods pilot study. *Med Teach*. 2017;39(10):1074-1083. doi:10.1080/0142159X.2017.1353070
26. Ginsburg S, Kogan JR, Gingerich A, Lynch M, Watling CJ. Taken out of context: hazards in the interpretation of written assessment comments. *Acad Med*. 2020;95(7):1082-1088. doi:10.1097/ACM.0000000000003047
27. Englander R, Cameron T, Ballard AJ, Dodge J, Bull J, Aschenbrener CA. Toward a common taxonomy of competency domains for the health professions and

competencies for physicians. *Acad Med*. 2013;88(8):1088-1094. doi:10.1097/ACM.0b013e31829a3b2b
28. Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE. Construct validity of the mini-clinical evaluation exercise (miniCEX). *Acad Med*. 2003;78(8):826-830.
29. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24(1):74-79.
30. Monteiro S, Sullivan GM, Chan TM. Generalizability theory made simple(r): an introductory primer to G-studies. *J Grad Med Educ*. 2019;11(4):365-370. doi:10.4300/JGME-D-19-00464.1
31. Murphy MJ, De A Seneviratne R, Remers OJ, Davis MH. "Hawks" and "doves": effect of feedback on grades awarded by supervisors of student selected components. *Med Teach*. 2009;31(10):e484-e488. doi:10.3109/01421590903258670