2024, Volume 56, Issue 8, 471-475, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Imprecise Clinical Assessments and Inaccurate Grades: Family Medicine Clerkship Director Perspectives

Arindam Sarkar, MD[a]; Joel J. Heidelbaugh, MD[b]; Gage Hallbauer, MD[a]; Nital P. Appelbaum, PhD[c]

AUTHOR AFFILIATIONS:

[a] Department of Family and Community Medicine, Baylor College of Medicine, Houston, TX

[b] Department of Family Medicine, University of Michigan Medical School, Ann Arbor, MI

[c] Department of Education, Innovation and Technology, Baylor College of Medicine, Houston, TX

HOW TO CITE: Sarkar A, Heidelbaugh JJ, Hallbauer G, Appelbaum NP. Imprecise Clinical Assessments and Inaccurate Grades: Family Medicine Clerkship Director Perspectives. *Fam Med.* 2024;56(8):471-475.
doi: 10.22454/FamMed.2024.819598

PUBLISHED: 13 June 2024

KEYWORDS: assessment, grading, undergraduate medical education

ABSTRACT

Background and Objectives: As application to residency programs becomes increasingly competitive, educational leaders face growing student concern about imprecise clinical assessments and clerkship grades.

Methods: As part of a large annual survey of family medicine clerkship directors (FMCDs), 10 questions were disseminated in May 2023 about perceived levels of imprecise assessments by faculty. We aimed to determine to what extent respondents felt their institution's evaluation system propagated inaccurate grading.

Results: A total of 52% of 169 FMCDs responded to the survey. Of these, 7% of respondents were completely confident that their preceptors would give two students of identical competence the same clinical evaluation rating. FMCDs estimated that an average of 38% of their preceptors inaccurately rate student performance. Most clerkships use an Honors/High Pass/Pass/Fail grading system. We found that 51% of FMCDs prefer to use a different grading paradigm than they currently use. We asked FMCDs to estimate the percentage of students that expressed concern over inaccurate preceptor ratings. Grading systems with more tiers were associated with a higher percentage of concerned students.

Conclusions: Clerkship grades are widely used by residency program directors to classify and differentiate student applicants. We identified a significant concern from FMCDs that clinical evaluation ratings can vary greatly. Given the high stakes and perceived inaccuracy of clerkship grading, we recommend continued investigation into the appropriate weighing and usage of clinical evaluations. Continued exploration is recommended to develop grading paradigms centered on criterion-based assessment.

INTRODUCTION

Medical school core clerkship grades are major influencers for residency program directors (PDs) in the selection of student applicants.[1–3] With the United States Medical Licensing Examination Step 1 now a Pass/Fail exam, clerkship grades have further risen in significance for PDs.[4] Clinical evaluations by faculty, often in the form of subjective assessments, are heavily weighted in clerkship grading rubrics.[5] Studies are mixed whether these clinical ratings are reflective of actual knowledge when compared to written and oral exams.[6–8] Clinical evaluations can be influenced by preceptors' habits, comparisons to other students, inflation pressures, and students' personalities.[9]

Significant variation in grading within and among medical schools has led to growing student concern nationwide about unfair grading from imprecise and inaccurate clinical ratings.[10] One measure of student concern, the grade challenge rate, has been described in other clerkships as ranging from 4.5% to 8.0% of students in a given year.[11,12] Family medicine clerkships may be especially subject to student concerns about fairness, given the diversity of academic, community, and volunteer faculty who perform clinical assessments. Grade inflation has been identified as a particular challenge for family medicine clerkship directors (FMCDs).[13]

No recent studies have reported on the prevalence of imprecise clerkship raters or FMCD's satisfaction with their current grading systems. Our study aimed to reveal associations among FMCD's perceived prevalence of improper faculty ratings, confidence in grading systems, length of tenure, and comfort level with addressing student dissatisfaction. We hypothesized that many FMCDs share concern regarding imprecise faculty ratings.

METHODS

Survey Administration and Development

Data were gathered and analyzed as part of the 2023 Council of Academic Family Medicine (CAFM) Educational Research Alliance (CERA) survey of FMCDs. CAFM is a joint initiative of four major academic family medicine organizations: the Society of Teachers of Family Medicine, the North American Primary Care Research Group, the Association of Departments of Family Medicine, and the Association of Family Medicine Residency Directors. The general methodology of the annual questionnaire, termed an omnibus survey, has been previously described.[14]

CERA distributed the survey via email to 169 (154 US and 15 Canadian) FMCDs between May 2023 and June 2023. The specific methodology for the 2023 survey has been described in detail.[15] The invitation email included an explanation and link to the online survey, which was conducted via SurveyMonkey (SurveyMonkey Inc). Nonrespondents received three requests to complete the survey.

Survey Questions

The survey featured a set of demographic questions to determine characteristics of the FMCDs and their clerkships. We submitted 10 additional closed-response survey items (Table 1) that assessed FMCD perceptions of imprecise grading, current and preferred grading systems, and comfort with correcting aberrant preceptor grading.

The first survey item established the perceived prevalence of imprecise preceptor ratings. FMCDs rated their confidence that preceptors would give two students of identical competence the same clinical evaluation rating (1=not at all confident, 5=completely confident). Respondents also rated the percentage of evaluators suspected of inappropriately rating student performance (0%-100%). FMCDs approximated the percentage of students that had expressed concern over their preceptor ratings within the last calendar year (0%-100%).

Next, FMCDs reported their current and preferred grading systems based on the options recognized by the Association of American Medical Colleges.[16] Subsequently, we asked for the percentage weight of clinical evaluations within grading rubrics (0%-100%) and whether FMCDs had recently adjusted that weight (multiple-choice item). Two items inquired whether grading systems accounted for outlier ratings (multiple-choice item) and which factors likely influenced preceptors' ratings (multiple-choice item). A final item assessed how comfortable FMCDs felt with remediating preceptors who inappropriately rate students (1=very uncomfortable, 5=very comfortable).

Analyses

We analyzed anonymous data with SPSS Version 26 (IBM) using descriptive statistics and Pearson correlations. We used Kruskal-Wallis H tests and Mann-Whitney U tests to determine the statistical significance of differences among respondent selections.

The American Academy of Family Physicians Institutional Review Board approved the study.

RESULTS

A total of 88 of 169 FMCDs (52%) responded to at least two survey items, and 85 FMCDs completed all 10 items. Of the respondents, 60% (n=53) were women and 68% were White (n=60). The average number of years spent in their current FMCD role was 6.64±5.08 (Table 2).

Prevalence of Imprecise Grading

We found that 53% (n=47) of FMCDs reported they were "not," "slightly," or "somewhat confident" their preceptors would give two students of identical competence the same clinical rating. Another 40% (n=35) were "fairly confident," and only 7% (n=6) of FMCDs were "completely confident" on this item.

FMCDs estimated that an average of 38% of their preceptors inaccurately rated student performance. Almost all respondents (n=86, 99%) perceived that at least 5% of their preceptors were overrating or underrating students. FMCDs estimated that 8% (SD=10%; range: 0-50%) of students expressed concern over their clinical ratings within the last year. We found no relationship between the number of years as FMCD and prevalence of perceived inaccurate assessors (Table 3). Similarly, years of experience as a FMCD was not correlated with estimated percentage of dissatisfied students.

Causes of Imprecise Grading

Inadequate faculty development (n=56, 64%), pressure to bestow high grades (n=55, 63%), and a fear of consequences to the preceptor (n=52, 59%) were selected as likely influencers of imprecise ratings. Of respondents, 14% (n=12) felt that none of these factors influenced their preceptors. Respondents who selected any of the potential above-mentioned influencers on Item 9 reported a higher perceived prevalence of imprecise grading on Item 2, compared to FMCDs who selected no influencers (P=.023, r=0.24; Table 3). Notably, the same CDs who believed that any of the three factors impacted their faculty estimated a higher percentage of students expressing concern over their clinical ratings (P<.001, r=0.38) compared to respondents denying the influences (Table 3).

Grading Systems

The average weight of clinical evaluations in clerkship grading rubrics was 57% (SD=21%; range: 0&-100%). Most FMCDs (n=72, 84%) had not decreased the percentage weight of clinical evaluations within their rubrics over the last 3 years. The majority (n=62, 70%) of FMCDs reported that their grading systems do not automatically account for outlier preceptor ratings.

Of respondents, 57% (n=50) used an Honors/High Pass/-Pass/Fail grading system (Table 4). We identified that 51% (n=44) of FMCDs preferred a different grading system than what they currently used. The two grading systems that were most often selected were Pass/Fail (n=29, 33%) and Honors/Pass/Fail (n=27, 31%).

TABLE 1. Ten Questions Posed to Clerkship Directors

1. Overall, how confident do you feel that all clerkship preceptors would give two students of identical competence the same clinical evaluation rating? (1=not at all confident, 5=completely confident)
2. What percentage of your clerkship preceptors do you suspect either overrate or underrate student performance on clinical evaluations? (0%-100%)
3. In 2022, what percentage of FM clerkship students expressed concern over their preceptor clinical evaluation ratings? (0%-100%)
4. In 2022, what grading tiers were used for your clerkship? (Pass/Fail, Honors/Pass/Fail, Honors/High Pass/Pass/Fail, Numerical Grade, Letter Grade, another category that is not listed)
5. What are your personally preferred grading tiers for your clerkship? (Pass/Fail, Honors/Pass/Fail, Honors/High Pass/Pass/Fail, Numerical Grade, Letter Grade, another category that is not listed)
6. In 2022, what percentage of your overall FM clerkship grade was determined by preceptors' clinical evaluation scores? (0%-100%)
7. Within the last 3 years, has the percentage weight of preceptors' clinical evaluations changed within your clerkship grading rubric? (No, Yes–% weight increased, Yes–% weight decreased by 1%-5%, Yes–% weight decreased by 6%-10%, Yes–% weight decreased by 10+%)
8. Does your current grading process account for outlier preceptor ratings (eg, eliminate highest/lowest ratings, use of educator bias reports, or perform statistical corrections)? (No, my clerkship grading process does not account for outlier preceptor ratings nor do I ask preceptors to modify their ratings; No, my clerkship grading process does not account for outlier preceptor ratings but I can ask individual preceptors to modify their ratings; Yes, my clerkship grading process accounts for outlier preceptor ratings)
9. Which of the following factors do you believe influences preceptor evaluations for your clerkship? (Pressure to bestow high grades, Inadequate faculty development, Fear of negative consequence to the preceptor (eg, student complaint, decrease in popularity among students, student retaliation, decrease in student rating of faculty), Combination of factors, None)
10. Please indicate your level of comfort in remediating preceptors who inappropriately rate students' clinical evaluation performance. (1=very uncomfortable, 5=very comfortable)

TABLE 2. Characteristics of Family Medicine Clerkship Director Respondents (N=88)

Characteristics	n (%)
Gender	
Female	53 (60%)
Male	34 (39%)
Missing/unknown	1 (1%)
Race/ethnicity	
Asian	15 (17%)
Black or African American	3 (3%)
White	60 (68%)
Hispanic or Latino/a	1 (1%)
Native American/Indigenous	0 (0%)
Missing/unknown	1 (1%)
Multiracial	1 (1%)
Average number of years in current CD role (Mean ± SD)	6.64 ± 5.08

Abbreviation: SD, standard deviation; CD, clerkship director

FMCDs with four grading tiers estimated a higher percentage of inaccurate clinical evaluations compared to FMCDs with three tiers. The estimated percentage of inaccurate clinical rating varied significantly based on the school's grading system (P<.001; Table 3). The estimated percentage of students dissatisfied with their preceptor ratings also varied significantly based on the school's grading system (P=.003). Namely, FMCDs using four grading tiers estimated more student complaints about grading compared to FMCDs using two or three tiers.

Comfort With Correcting Aberrant Preceptor Ratings

Of FMCD respondents, 55% (n=48) were either "somewhat" or "very comfortable" addressing and correcting preceptors who inappropriately rated students' clinical performance. We found no difference in comfort with remediating preceptors based on years of experience as an FMCD (Table 3).

DISCUSSION

One-half of responding FMCDs were not confident that their preceptors would give two students of identical competence the same clinical evaluation rating. Inadequate faculty development was most often selected by respondents as a contributor to improper ratings. We postulate that many FMCDs feel that adequate and effective faculty development is limited by the intuitiveness and specificity of their institutional assessment tool.

The estimated percentage of students challenging their clinical ratings was consistent with other clerkships reported in

TABLE 3. Group Comparison Tests

Groups	Confidence that preceptors would give two students of identical competence the same rating		Comfort in remediating preceptors who inappropriately rate students' clinical evaluation performance	
High/low years as CD	No group difference, $P>.05$		No group difference, $P>.05$	
Groups	Percentage of preceptors suspected over/underrating student performance on clinical evaluations		Percentage of students expressed concern over their preceptor clinical evaluation ratings	
	Test statistic	Pairwise comparisons and effect size	Test statistic	Pairwise comparisons and effect size
Evaluations influenced/Evaluations not influenced	z=2.27 P=.023 r=0.24	–	z=3.57 P<.001 r=0.38	–
Multiple grading tiers	H(2)=14.01 P<.001	Honors/Pass/Fail vs Honors/High Pass/Pass/Fail P=.002, r=0.42	H(2)=11.68 P=.003	Pass/Fail vs Honors/High Pass/Pass/Fail P=.007, r=0.38 Honors/Pass/Fail vsHonors/High Pass/Pass/Fail P=.018, r=0.35

Note: Analyses included Kruskal-Wallis and Mann-Whitney U tests, multiple comparisons corrections, and effect sizes (r).
Abbreviation: CD, clerkship director

TABLE 4. Preferred and Actual Grading System

Grading type	Actual grading	Preferred grading
Pass/Fail	12 (14%)	29 (33%)
Honors/Pass/Fail	13 (15%)	27 (31%)
Honors/High Pass/Pass/Fail	50 (57%)	23 (27%)
Numerical Grade	0 (0%)	1 (1%)
Letter Grade	9 (10%)	5 (6%)
Another category that is not listed	3 (3%)	1 (1%)
No response	1 (1%)	2 (2%)

the literature.[12] Despite FMCD awareness of inaccurate ratings, the percentage weight of clinical evaluations in clerkship grading has remained high. More grading tiers correlated with higher estimates of imprecise graders and higher estimates of student dissatisfaction. Nonetheless, FMCDs remained mixed in preferring tiered grading systems versus a Pass/Fail system. Possible reasons for preferring tiered grading may relate to residency stakeholders and considerations external to the clerkship.

LIMITATIONS

This study gathered perceptions of issues facing clerkships, students, and preceptors. While respondents could provide firsthand views on many survey items, several items required FMCDs to approximate information secondhand.

This 10-question survey explored only upstream causes and downstream effects of aberrant preceptor clerkship grading. We did not differentiate the prevalence of over- versus underrating of student performance by preceptors. Open-ended questions and matrix formats were not permitted in this survey. Narrative input on strategies to address imprecise grading was not gathered. While grade inflation and fear of retribution may be important factors in clinical assessments, these factors were not specifically measured for impact or influence.

Next steps

Interestingly, FMCDs with more grade options reported a higher frequency of student complaints. While Pass/Fail grading systems may reduce the harms of competitive stress and evaluator bias, a concurrent need exists for multidimensional assessments of students' strengths and weaknesses. The growing popularity of competency-based medical education encourages the use of directly observed assessments with behaviorally anchored rubrics. Criterion-driven tools, such as workplace-based assessments may reduce student perceptions of subjectivity; however, the de-emphasis on norm-based assessments may increase PDs' difficulty in ranking and stratifying students in effective and equitable ways.

CONCLUSIONS

Essentially all responding FMCDs perceived some level of improper faculty ratings in their courses. Despite clinical evaluation scores being the most heavily weighted component of their clerkship, a minority of FMCDs reported a systematic mechanism within their grade calculation process to account for outlier clinical ratings. In an effort to lessen the impact of inaccurate grading, medical schools in the United States and Canada should continue to explore ways to reduce subjectivity within their clinical assessment tools. Given the inherent challenges to delivering faculty development to busy clinical

preceptors, clinical rating anchors based on directly observed behaviors may improve rating precision.

REFERENCES

1. Green M, Jones P, Thomas JX. Selection criteria for residency: results of a national program directors survey. *Acad Med.* 2009;84(3):362-367.
2. Bird JB, Friedman KA, Arayssi T, Olvet DM, Conigliaro RL, Brenner JM. Review of the medical student performance evaluation: analysis of the end-users' perspective across the specialties. *Med Educ Online.* 2021;26(1):1876315.
3. National Resident Matching Program. Results of the 2021 NRMP Program Director Survey. 2021. .
4. Mcdonald JA, Lai CJ, Lin M, Sullivan O, Hauer PS, E K. There is a lot of change afoot": a qualitative study of faculty adaptation to elimination of tiered grades with increased emphasis on feedback in core clerkships. *Acad Med.* 2021;96(2):263-270.
5. Boatright D, Edje L, Gruppen, et al. Foundation Conference on Ensuring Fairness in Medical Education Assessment: conference recommendations report. *Acad Med.* 2023;98(8S):S3-S15.
6. Bullock JL, Lai CJ, Lockspeiser T. In pursuit of honors: a multi-institutional study of students' perceptions of clerkship evaluation and grading. *Acad Med.* 2019;94(11S):48-56.
7. Russo RA, Raml DM, Kerlek AJ, Klapheke M, Martin KB, Rakofsky JJ. Bias in medical school clerkship grading: is it time for a change?. *Acad Psychiatry.* 2023;47:428-431. .
8. Valentine N, Durning SJ, Shanahan EM, Vleuten CVD, Schuwirth L. The pursuit of fairness in assessment: looking beyond the objective. *Med Teach.* 2022;44(4):353-359.
9. Khan MA, Malviya M, English K. Medical student personality traits and clinical grades in the internal medicine clerkship. *Med Sci Educ.* 2021;31(2):637-645.
10. Smith JF, Piemonte NM. The problematic persistence of tiered grading in medical school. *Teach Learn Med.* 2023;35(4):467-476.
11. Packer CD, Duca NS, Dhaliwal G. Grade appeals in the internal medicine clerkship: a national survey and recommendations for improvement. *Am J Med.* 2021;134(6):817-822.
12. Thomas LA, Milburn N, Kay A, Hatch E. Factors associated with grade appeals: a survey of psychiatry clerkship directors. *Acad Psychiatry.* 2018;42(3):354-356.
13. Schiel KZ, Everard KM. Grade inflation in the family medicine clerkship. *Fam Med.* 2019;51(10):806-810.
14. Seehusen DA, Mainous AG, Iii, Chessman AW. Creating a centralized infrastructure to facilitate medical education research. *Ann Fam Med.* 2018;16(3):257-260.
15. Kost A, Moore MA, Ho T, Biggs R. Protocol for the 2023 CERA Clerkship Director Survey. *PRiMER.* 2023;7:30.
16. Grading systems used in medical school programs. *Association of American Medical Colleges.* 2009. https://www.aamc.org/data-reports/curriculum-reports/data/grading-systems-used-medical-school-programs.