Facilitator: Dorene Balmer, PhD
Discussant: Jennifer Kogan, MD

In Pursuit of Honors: A Multi-Institutional Study of Students' Perceptions of Clerkship Evaluation and Grading

Justin L. Bullock, MPH, Cindy J. Lai, MD, Tai Lockspeiser, MD, MHPE, Patricia S. O'Sullivan, EdD, Paul Aronowitz, MD, Deborah Dellmore, MD, Cha-Chi Fung, PhD, Christopher Knight, MD, and Karen E. Hauer, MD, PhD

Abstract

Purpose

To examine medical students' perceptions of the fairness and accuracy of core clerkship assessment, the clerkship learning environment, and contributors to students' achievement.

Method

Fourth-year medical students at 6 institutions completed a survey in 2018 assessing perceptions of the fairness and accuracy of clerkship evaluation and grading, the learning environment including clerkship goal structures (mastery- or performance-oriented), racial/ethnic stereotype threat, and student performance (honors earned). Factor analysis of 5-point Likert items (1 = strongly disagree, 5 = strongly agree) provided scale scores of perceptions. Using multivariable regression, investigators examined predictors of honors earned. Qualitative content analysis of responses to an open-ended question yielded students' recommendations to improve clerkship grading.

Results

Overall response rate was 71.1% (666/937). Students believed that being liked and particular supervisors most influenced final grades. Only 44.4% agreed that grading was fair. Students felt the clerkship learning environment promoted both mastery and performance avoidance behaviors (88.0% and 85.6%, respectively). Students from backgrounds underrepresented in medicine were more likely to experience stereotype threat vulnerability (55.7% vs 10.9%, $P < .0005$). Honors earned was positively associated with perceived accuracy of grading and interest in competitive specialties while negatively associated with stereotype threat. Students recommended strategies to improve clerkship grading: eliminating honors, training evaluators, and rewarding improvement on clerkships.

Conclusions

Participants had concerns around the fairness and accuracy of clerkship evaluation and grading and potential bias. Students expressed a need to redefine the culture of assessment on core clerkships to create more favorable learning environments for all students.

Preparing for clinical practice requires students to acquire broad and rapidly expanding skills and knowledge.[1] Simultaneously, students face increasing competition for residency positions, particularly in certain specialties.[2,3] Together, these demands create a taxing clinical learning environment, which may adversely affect learners.[4] One significant contributor to student stress is clerkship grading.[5,6] Grades provide important feedback to students and medical schools, and residency programs rely on core clerkship grades in resident selection.[7–9] Grade assignments are typically informed by examination scores and summative evaluations from supervising faculty and residents. Still, students and educators alike question the fairness and accuracy of grades.[4] Drawing from the educator's adage "assessment drives learning," negative perceptions of the current assessment system may adversely affect students' motivation, learning behaviors, and performance.[10]

Students' concerns around clerkship evaluations and grading may arise from a variety of factors. Supervisors variably interpret assessment scales and may lack a shared mental model of top performance.[11–13] Students can feel uncertain about what supervisors value when evaluating them.[14] A fair assessment system requires sufficient opportunities for students to learn and demonstrate learning, uses transparent criteria for evaluation and grading, and is equitable.[15,16] One study at a single medical school found that only 38% of students felt that clerkship evaluation was fair.[17] Students may doubt the accuracy of their evaluations because supervisors evaluate trainees on competencies despite infrequent direct observation of those trainees.[18,19] Bias also threatens accuracy and raises skepticism around grades. Students from racial or ethnic groups underrepresented in medicine (UIM) are less likely to earn top grades and honor society selection.[20–22]

All students can be susceptible to influences of the clerkship environment on their learning. A mastery-oriented environment fosters adaptive approaches to learning in which students seek challenges and thrive when facing obstacles.[23]

Please see the end of this article for information about the authors.

Correspondence should be addressed to Karen E. Hauer, Office of Medical Education, University of California, San Francisco, 533 Parnassus Ave., U80, Box 0710, San Francisco, CA 94143; telephone: (415) 502-5475; email: karen.hauer@ucsf.edu.

Acad Med. 2019;94:S48–S56.
First published online July 30, 2019
doi: 10.1097/ACM.0000000000002905
Copyright © 2019 by the Association of American Medical Colleges

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A720.

Conversely, performance-oriented environments include "performance approach," which rewards students for performing tasks that they know will make them appear competent, and "performance avoid," which encourages students to avoid challenging situations that could make them appear incompetent. The transition from a more mastery-oriented pass/fail preclinical learning environment to a more performance-oriented tiered grading clinical learning environment may cause students to deemphasize mastery-oriented behaviors and negatively affect learning.[24] A performance-oriented learning culture can decrease students' retention of information and satisfaction and increase burnout.[23,25]

Grading disparities between UIM and non-UIM students prompt consideration of other forces in the clerkship learning environment, beyond evaluator bias, which may uniquely contribute to poorer UIM student performance.[21,26] When vulnerable members of stigmatized groups (e.g., students from races/ethnicities typically UIM) worry that they will conform to lower expectations for their group, they experience stereotype threat. Stereotype threat exacerbates group differences in performance by increasing cognitive load and inhibiting the display of acquired skills and competencies.[27–29] While stereotype threats relating to race, gender, and age have been widely explored, a dearth of literature examines effects of stereotype threat amongst medical students.[28–32]

We designed this study to (1) examine students' perceptions of the fairness and accuracy of clerkship evaluation and grading, (2) examine students' perceptions of the clerkship learning environment, and (3) assess the relationship between these perceptions and students' achievement.

Method

Design

This is a multi-institutional, cross-sectional survey study.

Setting

Study institutions were a convenience sample of 6 U.S. schools in the Western Group on Educational Affairs, representing diverse western geographical locations and public/private status (Table 1). No invited schools declined participation. All 6 institutional review boards approved the study. All schools required students to complete family medicine, internal medicine, obstetrics–gynecology, pediatrics, psychiatry, and surgery clerkships (see Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/A720). Some had additional required clerkships. In this study, "honors" refers to the highest clerkship grade achievable at

Table 1

Demographic Data for Fourth-Year Medical Student Survey Respondents at 6 U.S. Medical Schools in 2018

	School[a] #1	School #2	School #3	School #4	School #5	School #6	Overall	*P* value[b]
Response rate (completed surveys) (%)[c]	81/89 (91.0)	127/168 (75.6)	132/170 (77.6)	148/237 (62.4)	111/185 (60.0)	67/88 (76.1)	666/937 (71.1)	—
Mean age, years (SD)	27.4 (2.5)	28.0 (3.1)	27.7 (2.9)	27.5 (2.9)	26.0 (1.4)	28.3 (4.5)	27.5 (3.0)	< .0005
Female, no. (%)	50 (61.7)	64 (52.9)	61 (46.9)	82 (56.6)	59 (54.1)	41 (63.1)	357 (54.8)	.197
Underrepresented minority, no. (%)[d]	18 (22.5)	29 (23.4)	16 (12.3)	8 (5.6)	10 (9.4)	25 (39.7)	106 (16.4)	< .0005
Lesbian, gay, bisexual, transgender, or queer, no. (%)	6 (7.5)	26 (21.1)	13 (10.0)	15 (10.4)	15 (14.7)	12 (18.8)	87 (13.5)	.033
First-generation college student, no. (%)	38 (46.9)	26 (20.8)	22 (16.8)	33 (22.4)	9 (8.2)	17 (26.6)	145 (22.0)	< .0005
Applying into more competitive specialty, no. (%)[e]	7 (8.9)	19 (16.7)	14 (10.6)	14 (10.1)	28 (26.2)	8 (11.9)	90 (14.1)	.003
No. of core clerkships completed, mean (SD)	6.09 (0.94)	6.87 (1.38)	7.34 (1.12)	6.15 (0.72)	6.89 (0.65)	6.96 (1.07)	6.72 (1.11)	—
Fraction of clerkship grades that were honors (SD)	0.35 (0.30)	0.46 (0.26)	0.36 (0.27)	0.42 (0.30)	0.28 (0.23)	0.46 (0.32)	0.39 (0.29)	—

[a] Participating institutions (in alphabetical order): Keck School of Medicine of the University of Southern California; University of California, Davis; University of California, San Francisco (UCSF); University of Colorado; University of New Mexico, Albuquerque; and University of Washington.

[b] Chi-square *P* value, except for age (ANOVA).

[c] Number of respondents meeting inclusion criteria / (number of surveys distributed − number of respondents NOT meeting inclusion criteria).

[d] Underrepresented in medicine: students who self-identify as African American, Latino/Latina/Hispanic, or Native American/Alaskan Native/Native Hawaiian.

[e] A specialty was considered competitive if it met 2 of the following 3 criteria using 2018 National Resident Matching Program (NRMP) data: probability of matching ≤ 90%, median Step 1 score of matched applicants ≥ 240, median Step 2 CK (Clinical Knowledge) ≥ 250. Competitive specialties included dermatology, diagnostic radiology, neurological surgery, ophthalmology, orthopedic surgery, otolaryngology, plastic surgery, radiation oncology, and urology.

each school. Consistent with medical schools nationally, schools varied in the percentage of students allowed to receive honors, presence of longitudinal integrated clerkships, and method of grade assignments.[33]

Participating students

Eligible participants were all medical students at the end of the core clerkship year. At 5 schools, students received an individualized email link to an electronic survey platform (www.qualtrics.com), signed by the lead investigator of that school. School-specific rules required that the email invitation go to the sixth school's class listserv. Nonrespondents received up to 3 weekly reminders. The survey was active for 30 days after release. Upon completion, participants could submit their email address via an outside website to receive a $10 electronic gift card. After data collection, a data analyst not otherwise involved in the study removed identifying information and assigned participants random identification numbers. Surveys were excluded if the student did not complete the demographics section or completed fewer than 3 clerkships.

Theoretical model and survey development

We developed a survey following guidelines for survey development.[34] Two authors (J.L.B., K.E.H.) reviewed the literature to identify key theories, evidence, and gaps surrounding students' perceptions of clerkship grading. One school (University of California, San Francisco [UCSF]) held a student town hall on clerkship grading with medical school deans. Based on the literature review and town hall feedback, we developed a model of students' perceptions of the fairness and accuracy of clerkship assessment, student motivation and effort, perceptions of feedback, clerkship learning environment, and contributors to students' achievement outcomes (Figure 1). Using this model, we developed and pilot-tested survey items at 2 study schools (UCSF, University of Colorado School of Medicine) with 23 students who provided feedback in writing or in 1 of 4 focus groups. The final survey also included adapted questions from the Manual for the Patterns of Adaptive Learning Scales (PALS) and the Stereotype Vulnerability Scale (SVS).[28,35] We modified the PALS Mastery, Performance Approach, and Performance Avoid Classroom Goal Structure scales and SVS stereotype threat items to reference "clerkships." We eliminated 3 original SVS items because of double-negative wording that confused pilot students.

The final 106 survey items addressed participant demographics, self-reported number of honors earned, number of clerkships taken, intended specialty, perceived impact of various domains on their final grade (scored 0–10), and our hypothesized predictors: perceptions of grading (fairness, accuracy) and clerkship learning environment (motivation, stereotype threat). Predictor questions used a 5-point Likert scale (strongly disagree [1] to strongly agree [5]). One open-ended question solicited students' recommendations to improve grading (see Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/A720).

Factor analysis

We used principal components analysis for data reduction, treating Likert scale questions as continuous 1–5 variables for perceptions of fairness and accuracy

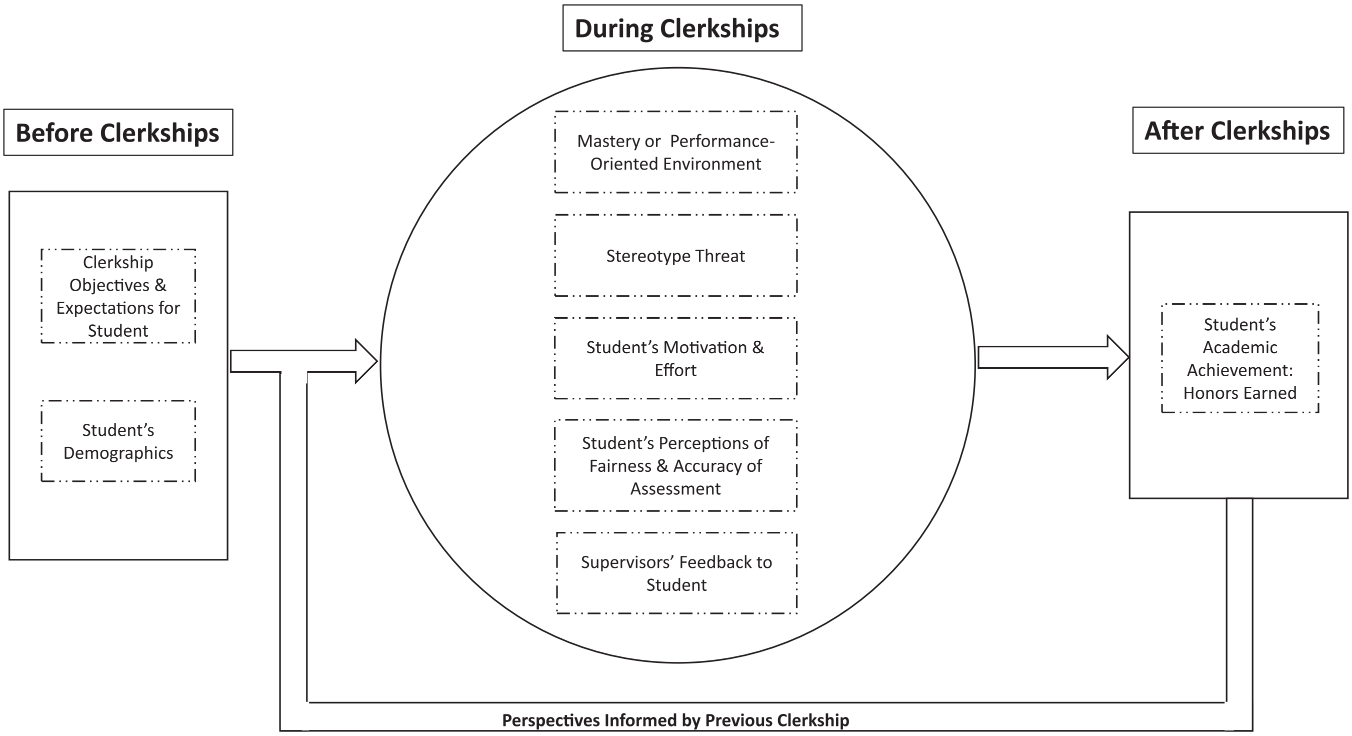
During Clerkships
Before Clerkships
After Clerkships
Clerkship Objectives & Expectations for Student
Student's Demographics
Mastery or Performance-Oriented Environment
Stereotype Threat
Student's Motivation & Effort
Student's Perceptions of Fairness & Accuracy of Assessment
Supervisors' Feedback to Student
Student's Academic Achievement: Honors Earned
Perspectives Informed by Previous Clerkship

Figure 1 Core clerkship student perceptions and outcomes model. Two authors conducted a review of medical education literature on evaluation and grading using the search terms education, medical, undergraduate, medical student, clinical clerkships, evaluation, grading, assessment, fairness, accuracy, motivation, mastery, performance, feedback, well-being, disparities, bias, learning environment, and stereotype threat. From this search, the authors constructed a theoretical model representing the major contributors to students' academic achievement on core clerkships.

of grading and clerkship learning environment. We used varimax rotation, retaining factors with an eigenvalue ≥ 1 and a maximum of 25 iterations before convergence. We used pairwise deletion for missing data. The Kaiser–Meyer–Olkin test was > 0.80, indicating sufficient correlation amongst items. Items were assigned to factors based on their largest loading. Because the PALS motivation scales and SVS were previously validated and still had high internal consistency with our minor modifications, they were not included in the principal component analysis.[28,35] For all factors, we calculated the Cronbach alpha coefficient and nonweighted mean score, retaining factors with Cronbach alpha > 0.6. Items were reverse-coded as needed so that all factor loadings were positive. For each retained factor, we calculated a scale score, treated as a continuous variable equal to the mean of the items comprising the factor. For scale scores, we categorized < 3 as "disagree," > 3 as "agree," and = 3 as "neutral." An SVS score > 3 indicated vulnerability to stereotype threat.

Statistical analysis

We calculated descriptive statistics for demographics. *t* Tests assessed differences in age. For all other subgroup comparisons, we used chi-square tests. To examine our first aim, we calculated descriptive statistics for students' perceptions of fairness and accuracy and students' experience in the clerkship learning environment. We used chi-square tests for subgroup comparisons of perceptions by gender and UIM status.

We used multivariable regression analysis to explore our second aim, the relationship between student demographics and perceptions and honors earned. To account for interschool differences in grading policies, we computed each student's standardized honors by calculating a *z* score using the fraction of clerkships honored, mean and standard deviation of the fraction of clerkships honored for that student's school. Hereafter, "honors earned" refers to each student's standardized honors value. We entered predictor variables in 2 blocks: student demographics and student perceptions (PCA-identified factors, PALS, SVS). We treated demographic variables as dichotomous except age, which was continuous. UIM students self-identified as African American, Latino, Latina, Hispanic, Native American, Alaskan Native, Native Hawaiian, or other Pacific Islander.[36] Using 2018 National Resident Matching Program data, competitive specialties were defined as meeting 2 of 3 criteria: probability of matching ≤ 90%, median Step 1 score of matched applicants ≥ 240, and median Step 2 CK (Clinical Knowledge) ≥ 250[37–40] (Table 1). We performed a Bonferroni correction to account for 16 comparisons in the regression, with a *P* value ≤ .003 deemed statistically significant.[41] We used IBM SPSS Statistics Version 23.0 for Windows (IBM, Armonk, New York) for analyses.

Qualitative analysis

Three authors (J.L.B., C.J.L., T.L.) analyzed comments using content analysis. Separately, each author inductively developed a codebook from a random sample of 50 comments. After discussion, we combined codes into a single codebook that we iteratively revised throughout the coding process. Using Microsoft Excel, 2 authors coded each comment independently and then reconciled discrepancies through discussion. Discussion of coding and attention to relationships among codes yielded key themes and subthemes. Code reconciliation naturally facilitated reflexivity as the coders included a senior medical student, clerkship director, and assessment committee director. We calculated the percentage of comments for which any portion of a student's comment applied to a given code.

Results

Overall, 972 students received survey invitations, 757 began the survey, and 701 completed it. Thirty-five students met exclusion criteria: 34 had completed fewer than 3 clerkships, and 1 reported earning more honors than clerkships taken. The final response rate was 666/937 (71.1%). Participants' mean age (SD) was 27.5 (3.0); 54.8% were women and 16.4% were UIM (Table 1). These percentages are similar to those in the national 2018 AAMC Medical School Graduate Questionnaire sample, among whom 49.1% were women and 15.5% were UIM.[42] Respondents had completed a mean (SD) of 6.7 (1.1) core clerkships. There were small, statistically significant differences across schools for mean age, percentage of UIM students, and percentage applying into competitive specialties (Table 1).

Perceived importance of domain on final grade

In response to the question: Considering the year as a whole, "in your experience, how important is each of the following in determining your final clerkship grade?" (see Supplemental Digital Appendix 3 at http://links.lww.com/ACADMED/A720), students scored "being liked" 8.7/10 (SD = 1.7), "particular attendings you work with" 8.7 (1.7), and "particular residents you work with" 8.5 (1.9) highest. They rated "improvement" 5.7 (2.7) and "rapport with patients and families" 6.0 (2.7) as least important.

Perceptions of grading

Our rotated PCA component matrix accounted for 64.9% of the total variance in our dataset and yielded 6 predictor factors (Table 2). Factors had high internal consistency (Cronbach alpha = 0.73–0.88). Students had low confidence in the fairness of grading, with only 44.4% of students agreeing that assessment was fair. Less than two-thirds of students felt that clerkship assessment was accurate or that feedback received was useful (60.8% and 61.7% agreed, respectively). Whereas 70.0% of students agreed that resident evaluation procedures were fair, only 41.7% agreed that attending evaluation procedures were fair.

One-third of students (33.6%) endorsed grading as biased. While more women perceived bias in evaluations than men (64.4% vs 25.2%, $P < .0005$), women also more commonly rated evaluations as accurate (69.2% vs 52.7%, $P < .0005$). There were no gender differences in perceptions of fairness of grading, feedback, or fairness of resident and attending evaluations. UIM students were more likely than non-UIM students to perceive bias in evaluations (48.1% vs 31.4%, $P = .0001$). Otherwise, UIM and non-UIM students' perceptions did not differ (see Supplemental Digital Appendix 4 at http://links.lww.com/ACADMED/A720).

Perceptions of the clerkship learning environment

Students overwhelmingly endorsed the clerkship learning environment to be both mastery- and performance-avoid-oriented (88.0% and 85.6%, respectively) (Table 2). Slightly fewer students

Table 2

Predictor Factors Identified Using Principal Components Analysis on Survey Items Answered by Students From 6 U.S. Medical Schools in 2018[a]

Factor (α coefficient)	No. of survey items	Description of higher score	Mean (SD)[b]	% Agree[c]
Predictor factors				
Grades are fair (á = 0.84)	7	Final clerkship grades reflect student performance based on clearly defined and fair criteria. Students understand the expectations upon which they are evaluated. Students can successfully appeal an unfair grade.	2.92 (0.85)	44.4
Evaluations are accurate (α = 0.87)	5	Evaluations of students are consistent and accurately reflect their clinical and interpersonal skills.	3.30 (0.98)	60.8
Students receive useful feedback (α = 0.80)	5	Feedback to students is useful and provides specific information on ways for students to improve.	3.26 (0.83)	61.7
Evaluations are biased (α = 0.88)	3	Students receive lower evaluations because of their intrinsic identity characteristics including gender, sexual orientation, race, and ethnicity.	2.93 (0.92)	33.6
Resident evaluation procedures are fair (α = 0.79)	3	Residents understand the assessment scale and observe students multiple times such that they know the student well enough to accurately evaluate them.	3.54 (0.98)	70.0
Attending evaluation procedures are fair (α = 0.76)	3	Attending physicians understand the assessment scale and observe students multiple times so that they know students well enough to evaluate them accurately.	2.89 (0.94)	41.7
Clerkship learning environment is mastery-oriented[d] (α = 0.82)	6	Mastery orientation entails a clerkship environment that values trying hard, improving, and mastering new material. In this environment, it is okay to make mistakes as long as student continues to learn.	4.03 (0.78)	88.0
Clerkship learning environment is performance-approach-oriented[d] (α = 0.73)	3	Performance approach entails a clerkship environment that values getting right answers, high scores on tests, and good grades.	3.54 (0.96)	68.9
Clerkship learning environment is performance-avoid-oriented[d] (α = 0.86)	5	Performance avoid entails a clerkship environment that values avoiding looking dumb. It is important for students to show that they are not bad at the work and don't make mistakes in front of others.	3.98 (0.81)	85.6
Student vulnerability to stereotype threat (α = 0.82)[e]	5	Evaluators expect some students to be less proficient because of their race or ethnicity. Students feel that they will receive biased evaluations because of their race or ethnicity.	2.36 (0.91)	18.3

[a]The authors performed a principal components analysis on items from the survey addressing students' perceptions of the fairness and accuracy of clerkship grading and the clerkship learning environment. We also included adapted Mastery, Performance Approach, and Performance Avoid Clerkship Goal Structure Scales and Stereotype Vulnerability Scale. All factors had a Cohen alpha value > 0.6.

[b]Items were coded 1 to 5, with 1 being minimally endorsing, 3 neutral, and 5 highly endorsing.

[c]Percentage of students with mean factor score > 3.

[d]Adapted from Patterns of Adaptive Learning Scales (PALS).[34]

[e]Adapted from original Stereotype Vulnerability Scale.[27]

endorsed clerkships as performance-approach-oriented (68.9%). There were no subgroup differences in perceptions of the mastery or performance orientation of clerkships by gender or UIM status.

Overall, 18.3% of student responses indicated vulnerability to stereotype threat based on race. Women and men perceived stereotype threat similarly. UIM students were much more likely than non-UIM students to indicate vulnerability to stereotype threat (55.7% vs 10.9%, $P < .0005$) (see Supplemental Digital Appendix 4 at http://links.lww.com/ACADMED/A720).

Honors earned multivariable regression analysis

Honors earned was positively associated with applying into a more competitive specialty (beta = 0.18, $P < .0005$) and perceiving evaluations as more accurate (beta = 0.29, $P < .0005$) (Table 3). Honors earned was negatively associated with stereotype threat (beta = −0.162, $P < .0005$). There were no significant associations between honors earned and perception of grading fairness, attending or resident evaluation procedures, or perceptions of mastery or performance environment of clerkships.

Qualitative analysis

Students' comments addressed 4 themes: grade assignment, evaluation process, bias causing differential grading, and learners' experience (Table 4). For grade assignment, many respondents recommended either reweighting components contributing to final grades or using pass/fail grading (29.3% of comments). Some recommended instituting competency-based assessment or using an entrustable professional activities system. In the evaluation process, students noted variability in assessors' knowledge of assessment and frameworks used to evaluate students.

Table 3
Multivariable Regression Assessing Predictors of Student Achievement[a]

Predictor variable	Standardized percent honors	
Adjusted R^2 value[b]	Full model = 19.6% Partial model = 5.9%	
	Standardized β	*P* value[c]
Age	−0.074	.050
Female	0.041	.289
Underrepresented minority[d]	−0.058	.161
Lesbian, gay, bisexual, or transgender	0.000	.997
First-generation college student	−0.042	.274
Applying into more competitive specialty[e]	**0.181**	**< .0005**
Fairness of grading	0.132	.022
Accuracy of evaluations	**0.290**	**< .0005**
Utility of feedback	−0.119	.013
Bias	−0.042	.296
Fair resident evaluation procedures	−0.040	.386
Fair attending evaluation procedures	0.051	.298
Mastery clerkship learning environment	−0.034	.437
Performance approach clerkship learning environment	0.058	.242
Performance avoid clerkship learning environment	−0.058	.239
Stereotype threat vulnerability	**−0.162**	**< .0005**

[a] Predictor variables were entered in 2 blocks: (1) student demographics and (2) scale scores (Patterns of Adaptive Learning Scales [PALS], Stereotype Vulnerability Scale [SVS], predictor factors).
[b] Partial model including only demographics variables. Full model included demographics and all predictor factors.
[c] To account for multiple comparisons, we used the Bonferroni correction with 16 comparisons per regression, with a final *P* value < 0.003 considered significant.
[d] Underrepresented in medicine indicates individuals who identify as African American, Latino/Latina/Hispanic, or Native American/Alaskan Native/Native Hawaiian.
[e] A specialty was considered competitive if it met 2 of the following 3 criteria using 2018 NRMP data: probability of matching ≤ 90%, median Step 1 score of matched applicants ≥ 240, median Step 2 CK (Clinical Knowledge) ≥ 250.

They recommended training evaluators on proper evaluation techniques (30.6%). To address biases causing differential grading, some advocated addressing evaluators' personal biases (19.2%) with implicit bias training or institutional systems to compare evaluators. To improve learners' experience, students wanted assessment to support learning through more regular and actionable feedback (14.4%), tracked over time so that improvement was valued and incorporated into final grades (11.6%).

Discussion

This multi-institutional study reveals low student confidence in the fairness of core clerkship evaluations and grading. More than half of UIM students endorsed stereotype threat vulnerability, a prevalence greater than 5 times that of non-UIM students. Perhaps unsurprisingly, students who were most successful in the current environment, defined by earning more honors, endorsed greater accuracy of evaluations, planned to apply in competitive specialties, and were less vulnerable to stereotype threat. Students' narrative comments supported their desire for changes to evaluation and grading.

Students' perceptions of grading have important implications for learning that should be addressed. Our results show that students perceive the strongest determinants of their grades as distinct from their clinical competence. Students who receive lower grades may attribute their grades to factors extrinsic to themselves such as an unfair system or variance of particular team members.[43,44] This scenario threatens self-efficacy and can negatively affect students' effort, behaviors, and future learning.[25,43] To address these challenges, our participants advocated for more evaluator training. While rater education is necessary for fair and accurate assessment of students' performance, there is inherent variability in the context and focus of particular patient encounters and evaluators themselves.[13,45] Rather than striving for perfect reliability among raters, a more appropriate goal would be to develop rigorous methods of collecting and synthesizing assessment data in a program of assessment.[46] However, adequate direct observation is also a necessary constituent of robust assessment. Our finding that students view residents' evaluations more favorably than attendings' may be explained by residents' greater direct contact with students working with patients. Increasing the number of observations from supervisors, in particular attending physicians, and exploring other mechanisms to improve students' experience with attending evaluators could improve students' perceptions of the fairness of evaluations.

Our data raise questions about whether the current assessment system promotes learning or performance.[47] Students felt that performance was highly valued, while improvement was minimally valued. The extrinsic motivation of an "honors" grade may promote a performance-oriented learning environment. In contrast, "assessment for learning" occurs when observations are used to both assess learning outcomes and provide timely, specific feedback, thereby transforming assessment into student learning.[9] This scenario cultivates mastery-oriented learners with improved long-term performance and enjoyment of learning.[23] Our participants' recommendations to redesign the clerkship assessment structure by eliminating tiered grades or changing to a competency-based approach could better promote a mastery mindset and lifelong learning.[48,49] Currently, the importance of grades for residency placement intensifies an already-high-pressure clerkship environment. Medical schools may hesitate to eliminate tiered clerkship grades because of their use during resident selection. While beyond the scope of our study, minimal data support that tiered clerkship grades effectively predict performance during residency.[50] Holistic review approaches by residency programs offer promise to reduce evaluation and grading pressures

Table 4

Students' Recommendations for Improving the Clerkship Evaluation and Grading Process, Based on Inductive Content Analysis of Written Comments From 396 Students From 6 U.S. Medical Schools in 2018

Themes and subthemes	Description	Supporting quotation(s)	No. (%) of comments n = 396
Grade assignment			
Redesign grading system	Recommendations were either for pass/fail clerkship grading or reweighting the components contributing to the grade. Some recommended instituting competency-based assessment or using an entrustable professional activities system.	"Make all clerkships pass/fail? The current grading scale is incredibly arbitrary, and I received similar grades and put in wildly different amounts of effort."	116 (29.3)
Evaluation process			
Transparency	Students felt that the grading process was not transparent to them, and they did not understand how faculty derived a grade.	"Have a systematic and universal way of grading clinical evaluations that is transparent to students."	44 (11.1)
Training evaluators	Students perceived variability in assessors' knowledge of assessment and inconsistencies in the framework used to determine clerkship grades. They commented on a need for more faculty development in this area and a need for improved rubrics to standardize grading.	"All evaluators should be formally trained in how the medical school's clerkship grading system works. They should be shown examples of good and bad medical student performance . . . to calibrate their grading scheme."	121 (30.6)
Effects of longitudinal relationships	To improve fairness, students desired more longitudinal relationships with evaluators and felt that the length of relationship with supervisors should be weighted for each evaluation.	"An evaluation from a provider that worked with you for a half day should be worth less than an evaluation from a provider that worked with you for a week."	40 (10.1)
Biases			
Clinical site	Students noted intersite variability in awarding of honors grades and in faculty awareness of expectations for student performance.	"There is certainly a discrepancy in grades at sites further away from the main campus hospital as they work with less students and may not understand the grading system."	48 (12.1)
Student personality	Students felt that personality qualities of an individual student influenced their residents' and attendings' evaluations of them. They felt that well-liked students received more favorable evaluations.	"I think clerkship grading is much harder for introverts. I don't know how to fix this because you cannot fix people's perceptions of how extroverted/confident you appear."	32 (8.1)
Evaluator	Students expressed concern about evaluator biases that influenced evaluations including implicit bias/racism and polar grading tendencies ("hawk" or "dove"). They recommended that supervisors undergo implicit bias training and that schools track and adjust for supervisor grading tendencies.	"Have all residents and attendings be trained in implicit biases and how they negatively affect trainees as well as patients, especially at a school that is not diverse in its class and the faculty are overwhelmingly white."	76 (19.2)
Evaluation	Students expressed concern around who does or does not fill out evaluations and endorsed infrequent direct observations by supervisors. They desired more observations, multiple evaluations, and requiring supervisors who had adequately observed them to fill out evaluations in a timely manner.	"I spent much more time with the fellow on my team and unfortunately, this person had moved to another service long before this attending completed my evaluation."	75 (18.9)
Learner's experience			
Feedback	Students expressed frustration that the written feedback used for their summative evaluations was inconsistent and lower than the in-person feedback they received. They wanted frequent, actionable feedback with improvement tracked over time.	"The clerkship where I feel I performed the best and where I received the strongest in-person feedback was the clerkship where my scores and final grade were the worst."	57 (14.4)
Growth	Students felt that grading created a maladaptive learning environment where students hesitate to ask questions or show ignorance because of grading repercussions. They also felt that improvement and responsiveness to feedback should be factored into grades.	"If this were a dance class, I could feel free asking my instructor which skills and moves to improve on to get an A in the course, but asking my attending what I need to do to get honors is very taboo and is either seen as manipulative or 'gunner.'"	46 (11.6)

for students and provide residencies useful information for selection.[51]

Stereotype threat vulnerability emerged as a significant negative predictor of performance, predominately affecting UIM students. UIM status was not a significant predictor of performance after controlling for stereotype threat vulnerability.

In addition to the documented grading biases facing UIM students, our findings support that stereotype threat may further undermine UIM students' academic achievement.[22,27] Despite being well described elsewhere, this phenomenon has not been explored amongst medical students. More work is needed to understand the scope and implications of stereotype threat in medical education and to design interventions to counteract it. Concrete strategies to mitigate the effects of stereotype threat include (1) introducing the concept of stereotype threat to the community, (2) engaging all community stakeholders to promote identity safety, and (3) increasing exposure to leaders of the stereotyped group.[52]

This study has limitations. Our results capture students' perspectives on clerkship grading; educators' opinions might differ. This cross-sectional survey does not show causation. Other unmeasured factors may contribute to student performance. Study schools are located in 1 U.S. region and may not generalize to other schools, although our study population was similar demographically to students nationally. We made small modifications to the PALS Classroom Goal Structures and SVS and assumed validity based on the original scales' validity in distinct populations. We did not collect performance data to correlate with survey responses, and students' specialty preferences may change over time. Finally, our qualitative results must be interpreted cautiously because students may have additional recommendations for clerkship grading that could have emerged with more questions, and not all students wrote comments.[53]

Our findings demonstrate that many medical students do not view evaluation and grading during core clerkships as fair, and they endorse an environment that encourages performance rather than rewards improvement. Negative perceptions of evaluation and grading are associated with decreased academic achievement. UIM students may face additional adverse pressures in the clerkship environment. A fair assessment system requires policies and procedures that promote equality and equity.[54] While many of the contributors hypothesized in our model (Figure 1) did not show associations with student performance, differential perceptions in these domains may have other effects such as changes in learning behaviors or student well-being.[55,56] These results support a need to redefine the culture of assessment on core clerkships to create learning environments that not only facilitate robust assessment but also enable learning for all students.

Acknowledgments: The authors wish to thank Alicia Fernandez, MD, and the University of California, San Francisco ESCape team for their thoughtful feedback on an earlier version of this article, as well as Robert Baraldi for his skillful statistical guidance.

Funding/Support: This project was supported by a Mini-Grant from the Association of American Medical Colleges Western Group on Educational Affairs.

Other disclosures: None reported.

Ethical approval: The institutional review board at each participating school approved this study.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the participating institutions.

J.L. Bullock is a first-year resident in internal medicine, Department of Medicine, University of California, San Francisco School of Medicine, San Francisco, California. The author was a fourth-year medical student at the time of writing.

C.J. Lai is director of internal medicine clerkships and professor, Department of Medicine, University of California, San Francisco School of Medicine, San Francisco, California.

T. Lockspeiser is director of the assessment/competency committee and associate professor, Department of Pediatrics, University of Colorado School of Medicine, Aurora, Colorado.

P.S. O'Sullivan is director of research and development in medical education and professor, Department of Medicine and Department of Surgery, University of California, San Francisco School of Medicine, San Francisco, California.

P. Aronowitz is clerkship director of internal medicine and professor, Department of Internal Medicine, University of California, Davis School of Medicine, Davis, California.

D. Dellmore is director of medical student education and associate professor, Department of Psychiatry and Behavioral Sciences, University of New Mexico School of Medicine, Albuquerque, New Mexico.

C.-C. Fung is assistant dean for medical education and associate professor, Keck School of Medicine of USC, Los Angeles, California.

C. Knight is associate clerkship director and associate professor, Division of General Internal Medicine, University of Washington School of Medicine, Seattle, Washington.

K.E. Hauer is associate dean for competency assessment and professional standards and professor, Department of Medicine, University of California, San Francisco School of Medicine, San Francisco, California.

References

1 Lucey CR. Medical education: Part of the problem and part of the solution. JAMA Intern Med. 2013;173:1639–1643.

2 Mullan F, Salsberg E, Weider K. Why a GME squeeze is unlikely. N Engl J Med. 2015;373:2397–2399.

3 Hayek S, Lane S, Fluck M, Hunsinger M, Blansfield J, Shabahang M. Ten year projections for US residency positions: Will there be enough positions to accommodate the growing number of U.S. medical school graduates? J Surg Educ. 2018;75:546–551.

4 Hauer KE, Lucey CR. Core clerkship grading: The illusion of objectivity. Acad Med. 2019;94:469–472.

5 Reed DA, Shanafelt TD, Satele DW, et al. Relationship of pass/fail grading and curriculum structure with well-being among preclinical medical students: A multi-institutional study. Acad Med. 2011;86:1367–1373.

6 Dyrbye L, Shanafelt T. A narrative review on burnout experienced by medical students and residents. Med Educ. 2016;50:132–149.

7 Green M, Jones P, Thomas JX Jr. Selection criteria for residency: Results of a national program directors survey. Acad Med. 2009;84:362–367.

8 Moss TJ, Deland EC, Maloney JV Jr. Selection of medical students for graduate training: Pass/fail versus grades. N Engl J Med. 1978;299:25–27.

9 Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. 2019;53:76–85.

10 Mennin SP, Kalishman S. Student assessment. Acad Med. 1998;73(9 suppl):S46–S54.

11 Durand RP, Levine JH, Lichtenstein LS, Fleming GA, Ross GR. Teachers' perceptions concerning the relative values of personal and clinical characteristics and their influence on the assignment of students' clinical grades. Med Educ. 1988;22:335–341.

12 de Jonge LPJWM, Timmerman AA, Govaerts MJB, et al. Stakeholder perspectives on workplace-based performance assessment: Towards a better understanding of assessor behaviour. Adv Health Sci Educ Theory Pract. 2017;22:1213–1243.

13 Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: Assessor cognition from three research perspectives. Med Educ. 2014;48:1055–1068.

14 Wimmers PF, Kanter SL, Splinter TA, Schmidt HG. Is clinical competence perceived differently for student daily performance on the wards versus clerkship grading? Adv Health Sci Educ Theory Pract. 2008;13:693–707.

15 Tierney RD. Fairness as a multifaceted quality in classroom assessment. Stud Educ Eval. 2014;43:55–69.

16 Colbert CY, French JC, Herring ME, Dannefer EF. Fairness: The hidden challenge for competency-based postgraduate medical education programs. Perspect Med Educ. 2017;6:347–355.

17 Duffield KE, Spencer JA. A survey of medical students' views about the purposes and fairness of assessment. Med Educ. 2002;36:879–886.

18 Schopper H, Rosenbaum M, Axelson R. 'I wish someone watched me interview:' Medical student insight into observation and feedback as a method for teaching communication skills during the clinical years. BMC Med Educ. 2016;16:286.

19 Howley LD, Wilson WG. Direct observation of students during clerkship rotations: A multiyear descriptive study. Acad Med. 2004;79:276–280.

20 Boatright D, Ross D, O'Connor P, Moore E, Nunez-Smith M. Racial disparities in medical student membership in the Alpha Omega Alpha Honor Society. JAMA Intern Med. 2017;177:659–665.

21 Teherani A, Hauer KE, Fernandez A, King TE Jr, 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;93:1286–1292.

22 Lee KB, Vaishnavi SN, Lau SK, Andriole DA, Jeffe DB. "Making the grade:" Noncognitive predictors of medical students' clinical clerkship grades. J Natl Med Assoc. 2007;99:1138–1150.

23 Dweck CS. Motivational processes affecting learning. Am Psychol. 1986;41:1040–1048.

24 Steinauer JE, O'Sullivan P, Preskill F, Ten Cate O, Teherani A. What makes "difficult patients" difficult for medical students? Acad Med. 2018;93:1359–1366.

25 Crooks TJ. The impact of classroom evaluation practices on students. Rev Educ Res. 1988;58:438–481.

26 Stegers-Jager KM, Steyerberg EW, Cohen-Schotanus J, Themmen AP. Ethnic disparities in undergraduate pre-clinical and clinical performance. Med Educ. 2012;46:575–585.

27 Spencer SJ, Logel C, Davies PG. Stereotype threat. Annu Rev Psychol. 2016;67:415–437.

28 Spencer SJ. The Effect of Stereotype Vulnerability on Women's Math Performance [doctoral thesis]. Ann Arbor, MI: University of Michigan; 1993.

29 Steele CM. A threat in the air. How stereotypes shape intellectual identity and performance. Am Psychol. 1997;52:613–629.

30 Schmader T, Johns M, Forbes C. An integrated process model of stereotype threat effects on performance. Psychol Rev. 2008;115:336–356.

31 Woolf K, Cave J, Greenhalgh T, Dacre J. Ethnic stereotypes and the underachievement of UK medical students from ethnic minorities: Qualitative study. BMJ. 2008;337:a1220.

32 Lamont RA, Swift HJ, Abrams D. A review and meta-analysis of age-based stereotype threat: Negative stereotypes, not facts, do the damage. Psychol Aging. 2015;30:180–193.

33 Alexander EK, Osman NY, Walling JL, Mitchell VG. Variation and imprecision of clerkship grading in U.S. medical schools. Acad Med. 2012;87:1070–1076.

34 Artino AR Jr, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE guide no. 87. Med Teach. 2014;36:463–474.

35 Midgley C, Maehr ML, Hruda LZ, et al. Manual for the Patterns of Adaptive Learning Scales. Ann Arbor, MI: University of Michigan; 2000. http://www.umich.edu/~pals/PALS%202000_V13Word97.pdf. Accessed July 18, 2019.

36 Association of American Medical Colleges. Underrepresented in medicine definition. https://www.aamc.org/initiatives/urm. Accessed July 18, 2019.

37 National Resident Matching Program. Charting outcomes in the match: U.S. allopathic seniors: Characteristics of U.S. allopathic seniors who matched to their preferred specialty in the 2018 main residency match. https://www.nrmp.org/wp-content/uploads/2018/06/Charting-Outcomes-in-the-Match-2018-Seniors.pdf. Published July 2018. Accessed July 18, 2019.

38 American Urology Association. 2018 urology residency match—Statistics. https://www.auanet.org/Documents/education/specialty-match/2018-Urology-Residency-Match-Statistics.pdf. Accessed July 18, 2019.

39 Association of University Professors of Ophthalmology. Ophthalmology residency match summary report 2018. https://www.sfmatch.org/PDFFilesDisplay/Ophthalmology_Residency_Stats_2018.pdf. Accessed July 18, 2019.

40 Jena AB, Arora VM, Hauer KE, et al. The prevalence and nature of postinterview communications between residency programs and applicants during the match. Acad Med. 2012;87:1434–1442.

41 Pedhazur EJ, Kerlinger FN. Multiple Regression in Behavioral Research: Explanation and Prediction. 2nd ed. New York, NY: Holt, Rinehart and Winston; 1982.

42 Association of American Medical Colleges. Medical school graduation questionnaire: 2018 all schools summary report. https://www.aamc.org/download/490454/data/2018gqallschoolssummaryreport.pdf. Accessed July 18, 2019.

43 Weiner B. Attribution theory, achievement motivation and the educational process. Rev Educ Res. 1972;42:203–215.

44 Graham S. A review of attribution theory in achievement contexts. Educ Psychol Rev. 1991;3:5–39.

45 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;18:325–341.

46 van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012;34:205–214.

47 Govaerts MJB, van der Vleuten CPM, Holmboe ES. Managing tensions in assessment: Moving beyond either-or thinking. Med Educ. 2019;53:64–75.

48 Pereira AG, Woods M, Olson APJ, van den Hoogenhof S, Duffy BL, Englander R. Criterion-based assessment in a norm-based world: How can we move past grades? Acad Med. 2018;93:560–564.

49 Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE guide no. 99. Med Teach. 2015;37:983–1002.

50 Lavin B, Pangaro L. Internship ratings as a validity outcome measure for an evaluation system to identify inadequate clerkship performance. Acad Med. 1998;73:998–1002.

51 Conrad SS, Addams AN, Young GH. Holistic review in medical school admissions and selection: A strategic, mission-driven response to shifting societal needs. Acad Med. 2016;91:1472–1474.

52 Burgess DJ, Joseph A, van Ryn M, Carnes M. Does stereotype threat affect women in academic medicine? Acad Med. 2012;87:506–512.

53 LaDonna KA, Taylor T, Lingard L. Why open-ended survey questions are unlikely to support rigorous qualitative insights. Acad Med. 2018;93:347–349.

54 Bierer SB, Colbert CY, Foshee CM, French JC, Pien LC. Last page: Tool for diagnosing gaps within a competency-based assessment system. Acad Med. 2018;93:512.

55 Pelaccia T, Viau R. Motivation in medical education. Med Teach. 2017;39:136–140.

56 Wasson LT, Cusmano A, Meli L, et al. Association between learning environment interventions and medical student well-being: A systematic review. JAMA. 2016;316:2237–2252.