

Teaching and Learning in Medicine

An International Journal


ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/htlm20>


Moving toward Mastery: Changes in Student Perceptions of Clerkship Assessment with Pass/Fail Grading and Enhanced Feedback

Justin L. Bullock, Lee Seligman, Cindy J. Lai, Patricia S. O'Sullivan & Karen E. Hauer


To cite this article: Justin L. Bullock, Lee Seligman, Cindy J. Lai, Patricia S. O'Sullivan & Karen E. Hauer (2021): Moving toward Mastery: Changes in Student Perceptions of Clerkship Assessment with Pass/Fail Grading and Enhanced Feedback, Teaching and Learning in Medicine, DOI: [10.1080/10401334.2021.1922285](https://doi.org/10.1080/10401334.2021.1922285)

To link to this article: <https://doi.org/10.1080/10401334.2021.1922285>

 View supplementary material 

 Published online: 20 May 2021.




 Submit your article to this journal 

 Article views: 243

 View related articles 

 View Crossmark data 

Moving toward Mastery: Changes in Student Perceptions of Clerkship Assessment with Pass/Fail Grading and Enhanced Feedback

Justin L. Bullock^a , Lee Seligman^b, Cindy J. Lai^a, Patricia S. O'Sullivan^{a,c} , and Karen E. Hauer^a 

^aDepartment of Medicine, University of California, San Francisco School of Medicine, San Francisco, California, USA; ^bDepartment of Medicine, Columbia University Irving Medical Center, New York-Presbyterian Hospital, New York, New York, USA; ^cDepartment of Surgery, University of California, San Francisco School of Medicine, San Francisco, California, USA

ABSTRACT

Problem: Clerkship grades contribute to a summative assessment culture in clerkships and can therefore interfere with students' learning. For example, by focusing on summative, tiered clerkship grades, students often discount accompanying feedback that could inform future learning. This case report seeks to explore whether an assessment system intervention which eliminated tiered grades and enhanced feedback was associated with changes in student perceptions of clerkship assessment and perceptions of the clinical learning environment. **Intervention:** In January 2019, our institution eliminated tiered clerkship grading (honors/pass/fail) for medical students during the core clerkship year and implemented pass/fail clerkship grading along with required twice weekly, work-based assessments for formative feedback. **Context:** In this single institution, cross-sectional survey study, we collected data from fourth-year medical students one year after an assessment system intervention. The intervention entailed changing from honors/pass/fail to pass/fail grading in all eight core clerkships and implementing an electronic system to record twice-weekly real-time formative work-based assessments. The survey queried student perceptions on the fairness and accuracy of grading and the clinical learning environment—including whether clerkships were mastery- or performance-oriented. We compared responses from students one year after the assessment intervention to those from the class one year before the intervention. Comparisons were made using unpaired, two-tailed t-tests or chi-squared tests as appropriate with Cohen's d for effect size estimation for score differences. Content analysis was used to analyze responses from two open-ended questions about feedback and grading. **Impact:** Survey response rates were similar before and after intervention (76% (127/168) vs. 72% (118/163), respectively) with no between-group differences in demographics. The after-intervention group showed statistically significant increases in the following factors: "grades are transparent and fair" (Cohen's d=0.80), "students receive useful feedback" (d=0.51), and "resident evaluation procedures are fair" (d=0.40). After-intervention respondents perceived the clerkship learning environment to be more mastery-oriented (d=0.52), less performance approach-oriented (d=0.63), and less performance avoid-oriented (d=0.49). There were no statistical differences in the factors "attending evaluation procedures are fair," "evaluations are accurate," "evaluations are biased," or "perception of stereotype threat." Open-ended questions revealed student recommendations to improve clerkship summary narratives, burden of work-based assessment, and in-person feedback. **Lessons Learned:** After an assessment system change to pass/fail grading with work-based assessments, we observed moderate to large improvements in student perceptions of clerkship grading and the mastery orientation of the learning environment. Our intervention did not improve perceptions around bias in assessment in clerkships. Other medical schools may consider similar interventions to begin to address student concerns with clerkship assessment and promote a more adaptive learning environment.

ARTICLE HISTORY

Received 17 November 2020
Revised 08 April 2021
Accepted 20 April 2021



KEYWORDS

Undergraduate medical education; pass fail grading; assessment; fairness; feedback

Introduction

The pervasive summative assessment culture in medical education has important effects on learning.^{1–4} Students and supervisors alike perceive formative assessment intended as "assessment for learning" to

instead be summative and "assessment of learning."^{1,3} Although essential for high-stakes decisions, summative assessment can interfere with the learning process. For example, by focusing on summative grades, students often discount accompanying feedback that

CONTACT Karen E. Hauer  karen.hauer@ucsf.edu  University of California, San Francisco, 533 Parnassus Ave, U80, Box 0710 San Francisco, CA 94143, USA

 Supplemental data for this article is available online at <https://doi.org/10.1080/10401334.2021.1922285>.

© 2021 Taylor & Francis Group, LLC

could inform future learning.⁵ In reality, most assessments lie along a continuum between purely summative or purely formative, and contextual factors influence how the stakes of an assessment are perceived.² Most American institutions assign tiered clerkship grades (e.g. honors/pass/fail or A/B/C/D), which are high-stakes and summative.⁶ Using grades to select students for medical school awards and residency program interviews promotes extrinsic motivation in students who seek to earn these accolades, perhaps at the expense of maximizing their learning.^{7–9} Many students question the fairness of grades, and or students from backgrounds typically underrepresented in medicine (UIM), concerns about bias in the assessment process compound concerns about fairness in grading.^{8,10,11} Together, these issues prompt consideration of how to optimize clerkship assessment to promote learning and create a healthier educational milieu.

Tiered clerkship grades can discourage a mastery mindset in learners and instead promote a performance mindset.^{10,12} Mastery-oriented learners learn for the sake of learning; they challenge themselves and persist despite setbacks.¹³ Performance approach-oriented learners engage in tasks that support the appearance of competence and seek to display already-mastered skills. Performance avoid-oriented learners avoid activities which could undermine their persona of proficiency, for instance, forgoing asking questions to avoid seeming incompetent.¹³ Though a performance approach mindset can lead to short-term achievement, performance-oriented behaviors do not optimize long-term development when compared to mastery-oriented behaviors, which better address learning gaps.¹⁴

Self Determination Theory (SDT) further elucidates student motivation for learning. According to SDT, learners move along a motivation spectrum from amotivation to extrinsic motivation to intrinsic motivation.^{15,16} Whereas extrinsically motivated individuals pursue activities for the purpose of achieving a reward or avoiding a negative outcome, intrinsically motivated individuals pursue activities because of inherent interest.¹⁶ “Autonomous motivation” encompasses intrinsic motivation and an adaptive form of extrinsic motivation in which the student accepts and endorses the expectations and rules made by others.¹⁷ Autonomously motivated learners regulate their own learning rather than yielding control of learning and the value placed on learning to others.^{18,19} Autonomous motivation is associated with higher academic performance and lifelong learning habits in both medical trainees and practicing physicians.^{18,20–22} Lifelong learning refers to self-initiated activities and

information-seeking skills with sustained motivation to learn as individuals recognize their own learning gaps and take action to self-improve.²² Extrinsically motivated assessment practices negatively impact intrinsic motivation²³ and may threaten the development of lifelong learning practices.

Pass/fail clerkship grading is one proposed solution to promote mastery-oriented behaviors and foster autonomous motivation.^{24–26} Some educators assume that traditional tiered grading is necessary for learner motivation.²⁷ Yet, grades constitute extrinsic motivators that may produce less desirable long-term learning outcomes, and medical students describe high intrinsic motivation to learn clinical medicine.²⁸ SDT also posits that all learners are naturally inclined to develop and will continue to do so, even in the absence of extrinsic motivators such as tiered grades.¹⁶ Pre-clinically, pass/fail grading has been widely accepted, and evidence demonstrates that it consistently enhances student well-being without compromising learning or performance in pre-clinical curricula.^{29,30}

Given institutional and national data highlighting pervasive concerns around tiered clerkship grading, our institution recently changed the third-year core clerkship assessment system by implementing pass/fail grading and augmenting feedback with required formative workplace-based assessments.^{8,11,31} Formative feedback was an important component of our assessment intervention: medical students commonly report not being observed by supervisors and not receiving regular or timely feedback necessary to guide self-improvement.^{32,33} Formalizing this feedback process simultaneously increased the amount of feedback for students and also provided an opportunity for supervisors to signpost to learners that they are receiving feedback. The consequences of these changes for students’ perceptions are unknown.

The purpose of this study was to examine whether pass/fail grading and enhanced formative feedback were associated with changes in student perceptions of the fairness and accuracy of clerkship grading and the orientation of the clerkship learning environment. We also aimed to explore student recommendations to maximize the fairness of the new assessment system and to improve the utility of feedback.

Methods

This single-institution, before-after cross-sectional survey study compares survey responses from students one year after implementation of an assessment system intervention to responses from students one year before the intervention. Our research team took part

in a previously published multi-institutional study on clerkship assessment; here we take our institution's subset of that data and compare it to new data collected after an assessment intervention.¹⁰ The University of California, San Francisco (UCSF) School of Medicine institutional review board deemed the study exempt (IRB #17-23328).

Intervention

UCSF changed the core clerkship assessment system by replacing tiered grades (honors/pass/fail) with pass/fail grading and work-based assessments (WBAs) for frequent feedback (Figure 1). Before the grading change, students received tiered clerkship grades in all required

used the same information as before the intervention to assign grades, except that interns no longer completed written evaluations. Students continued to receive summary evaluations. Additionally, students engaged in frequent low-stakes WBAs using a brief electronic observation tool for supervisors to record feedback to students. Students were required to collect WBAs twice weekly, but content from the WBAs was not used to determine clerkship summary evaluations or grades. In both time periods, the summary narratives were presented in the Medical Student Performance Evaluation (MSPE) verbatim. In the before- but not the after-period, clerkship grades were used to assign adjectives to designate standing within the class for the MSPE.

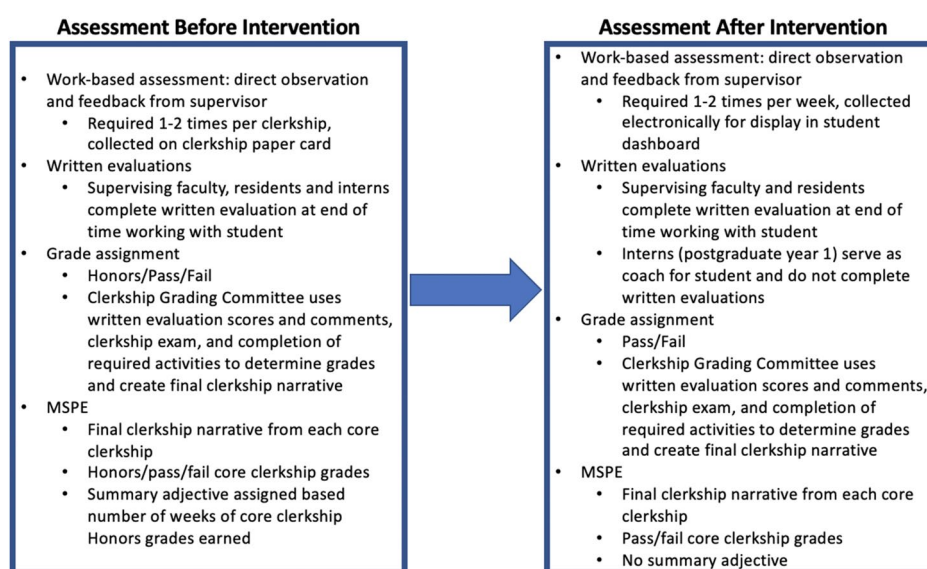


Figure 1. Description of assessment system intervention.

core clerkships (Anesthesia, Family Medicine, Medicine, Neurology, Obstetrics/Gynecology, Pediatrics, Psychiatry, Surgery), with a cap of 45% honors in each clerkship. Clerkship grading committees composed of the clerkship director and site directors determined students' grades based on written evaluations from supervising interns, residents, and attendings, a written examination (in all clerkships except Anesthesia), and completion of required clerkship activities.³⁴ Students received a summary evaluation written by the clerkship director for each rotation, which synthesized average numerical scores and narrative comments (which were not weighted quantitatively) from individual supervisor evaluations. The exam contributed 10–25% to the final grade in each clerkship.

After the grading change, students received a final clerkship grade of “pass” or “fail.” Grading committees

Time points

Summer 2018 (before group): We assessed clerkship students' perceptions after completing core clerkships under the honors/pass/fail grading system. **Spring 2020 (after group):** We assessed student perceptions of clerkship grading among the first class to complete core clerkships under the new pass/fail grading system. This class completed core clerkships earlier in the year than in the prior system and, as such, data collection moved to spring. Members of the fourth-year class in the 2018–2019 academic year were not included in this study as they completed clerkships while the intervention was introduced.

At both time points, students received an email invitation to complete a web-based survey (www.qualtrics.com) at least six weeks after completing their

final clerkship. This timing allowed students to receive grades from all core clerkships in compliance with Liaison Committee on Medical Education standards.³⁵ At survey completion, students were offered a \$10 gift card.

Survey

In previous work, we designed and provided validity evidence for a survey which tested a model of students' perceptions of clerkship assessment and grading, the clerkship learning environment, and the relationship of the environment to achievement outcomes.^{10,36} We used factor analysis to determine underlying latent constructs from survey items about perceptions of clerkship evaluation and grading.¹⁰ Six factors had eigenvalues greater than 1 and were therefore retained. The before-intervention survey also included items from the Manual for the Patterns of Adaptive Learning Scales and the Stereotype Vulnerability Scale.^{37,38} All factors showed good reliability (Cronbach α = 0.76–0.88).

The after-intervention survey contained 86 items. We eliminated the 20 items from the before-intervention survey which did not map to a retained factor and removed one item assessing honors earned. Retained survey items addressed student perceptions of fairness, accuracy, and the learning environment: “grades are fair and transparent” (7 items), “resident evaluation procedures are fair” (3), “attending evaluation procedures are fair” (3), “evaluations are accurate” (6), “students receive useful feedback” (5), “evaluations are biased” (3), orientation of the learning environment (14), vulnerability to stereotype threat (5), learning behaviors (17), contributors to grades (12), student demographics (9), and two free-response questions soliciting recommendations on how to improve clerkship assessment and feedback.^{37,38} Non-demographic survey items were scored on a Likert scale from 1 (strongly disagree) to 5 (strongly agree) except for “contributors to final grades,” for which students rated on a 0 to 10 scale how important each contributor was to determining their final grades. Survey items mapping to each factor are shown in Online Appendix 1.

Analysis

We calculated descriptive statistics for all variables. To assess differences between the two groups for the perceptions of assessment and learning environment factors, we calculated a scale score for each factor, treated as a continuous variable equal to the mean of items composing the factor. We used unpaired

two-tailed t-tests to compare before/after scale means. We used the Bonferroni correction to account for repeated comparisons.³⁹ With the correction factor, our p-value of significance was <0.005; as such we report 99.5% confidence intervals. To contextualize the magnitude of factor differences, we calculated Cohen's d using group means and the pooled standard deviation of our sample. Cohen's d = 0.2 is considered a “small” effect size, 0.5 “medium,” and 0.8 “large.”⁴⁰ To compare class demographics and perceived contributors to grades, we used unpaired, two-tailed t-tests or chi-squared tests as appropriate. To explore group differences in response to the intervention, we conducted a difference in difference analysis using linear regression. We performed one regression for each factor-group combination (e.g., change in stereotype threat for UIM and non-UIM respondents). The factor served as the dependent variable and the independent dichotomous variables included group (UIM, LGBTQ, or first generation students), intervention, and group-by-intervention interaction term.

Because our system intervention included both a WBA tool for feedback and pass/fail grading, one open-ended question addressed each. Three authors (J.L.B., L.S., C.J.L.) analyzed comments from the two open-ended questions from the after-group using content analysis.⁴¹ For each open-ended question, two of the three coding authors separately developed a codebook from half of the written comments. Using a combination of inductive (based on themes within the data) and deductive (based on our previous clerkship grading study)¹⁰ approaches, we created a single codebook that we iteratively revised throughout the coding process. Using Microsoft Excel, the authors independently coded each comment, later reconciling discrepancies through discussion, so that each response was coded by two authors. Discussion of coding and attention to relationships among codes yielded key themes and subthemes. Coding team members included a first-year resident (J.L.B.), senior medical student (L.S.), and internal medicine clerkship director (C.J.L.). We considered reflexivity through conversations about our various potential interpretations of comments and each researcher's perspective from having engaged in at least one previous study on clerkship grading.⁴² For the second open-ended question, we compared key themes from after-intervention to our previously published before-grading change study.¹⁰

Findings

After the grading change and feedback intervention, 72% (118/163) of invited students completed the

survey. This response rate was similar to the before-intervention group of 76% (127/168) (Table 1). There were no statistical differences between the two groups for age, gender, UIM status, LGBTQ, first generation, intent to apply into a competitive specialty, or number of core clerkships completed (Table 1).

Student perceptions

Fairness and accuracy of clerkship grading

The after-intervention group showed significant positive differences in three of six factors addressing fairness and accuracy: “Grades are fair and transparent” (2.62 before vs. 3.28 after, Cohen’s $d=0.80$), “students receive useful feedback” (3.09 vs. 3.50, $d=0.51$), and “resident evaluation procedures are fair” (3.06 vs. 3.46, $d=0.40$) (Table 2). There were no significant differences between groups in the factors “attending

evaluation procedures are fair,” “evaluations are accurate,” or “evaluations are biased.”

Learning orientation of the learning environment

After-intervention respondents perceived the clerkship learning environment to be more mastery-oriented (3.91 before vs. 4.27 after, $d=0.52$), less performance approach-oriented (3.48 vs. 2.86, $d=0.64$), and less performance avoid-oriented (4.18 vs. 3.80, $d=0.49$) compared to the before-intervention group. Perception of stereotype threat did not differ between groups (2.54 vs. 2.42, $d=0.17$). Visual representation of this data with stacking histograms is in Online Appendix 2. Difference in difference analysis of perceptions revealed that UIM, LGBTQ, and first-generation students’ responses changed similarly compared to their respective non-minoritized peers (Online Appendix 3). UIM students endorsed higher racial/ethnic

Table 1. Demographic characteristics of survey respondents for two different classes of fourth-year medical students at one United States medical school before and after an assessment system intervention for core clerkships.

	Before Grading Change (Comparison) ^a	After Grading Change (Intervention)	p-value
Response Rate (%)	127/168 (75.6%)	118/163 (72.4%)	0.51
Mean Age, years (SD)	28.0 (3.1)	27.6 (2.9)	0.21
Female, no. (%)	64 (50.4)	73 (61.9)	0.22
Under-represented in Medicine (UIM), ^b no. (%)	29 (22.8)	32 (27.1)	0.58
LGBTQ, no. (%)	26 (20.5)	25 (21.2)	0.97
First generation college student, no. (%)	26 (20.5)	28 (23.7)	0.68
Applying to competitive specialty, ^c no. (%)	19 (15.0)	17 (14.4)	0.80
No. of core clerkships completed (SD)	6.87 (1.38)	7.26 (1.13)	0.02 ^d

^aBefore-intervention demographics were previously published as a part of a multi-institutional study.¹⁰

^bUnderrepresented in medicine: students who self-identify as African American, Latino/Latina/Hispanic, or Native American/Alaskan Native/Native Hawaiian.

^cA specialty was considered competitive if it met two of the following three criteria using 2018 NRMP data: probability of matching $\leq 90\%$, median Step 1 score of matched applicants ≥ 240 , median Step 2 CK ≥ 250 . Competitive specialties included: dermatology, diagnostic radiology, neurological surgery, ophthalmology, orthopedic surgery, otolaryngology, plastic surgery, radiation oncology, and urology.

^dNot statistically significant due to use of Bonferroni correction with significant p-value of <0.005 .

Table 2. Comparison of mean scale scores for survey factors of student perceptions of the fairness and accuracy of clerkship assessment and the clerkship learning environment, before and after an assessment intervention at one medical school.

Survey Factor (Number of items, Cronbach alpha) ^a	Before Intervention Factor Mean [99.5% CI] ^b	After Intervention Factor Mean [99.5% CI]	p-value ^c	Cohen’s d ^d
Student Perception of the Fairness and Accuracy of Clerkship Assessment				
Grades are fair and transparent (n=7, $\alpha=0.84$)	2.62 [2.39, 2.85]	3.28 [3.09, 3.47]	<0.0005	0.80
Students receive useful feedback (n=5, $\alpha=0.80$)	3.09 [2.87, 3.31]	3.50 [3.31, 3.69]	<0.0005	0.51
Resident evaluation procedures are fair (n=5, $\alpha=0.79$)	3.06 [2.79, 3.33]	3.46 [3.23, 3.69]	0.002	0.40
Evaluations are biased (n=3, $\alpha=0.88$)	3.24 [3.04, 3.44]	3.48 [3.30, 3.67]	0.016	0.31
Evaluations are accurate (n=3, $\alpha=0.87$)	3.17 [2.91, 3.43]	3.35 [3.15, 3.56]	0.11	0.20
Attending evaluation procedures are fair (n=3, $\alpha=0.76$)	2.72 [2.45, 2.99]	2.85 [2.62, 3.08]	0.280	0.13
Student Perception of the Clerkship Learning Environment^{36,37}				
Clerkship learning environment is mastery-oriented (n=6, $\alpha=0.81$)	3.91 [3.71, 4.11]	4.27 [4.12, 4.41]	<0.0005	0.52
Clerkship learning environment is performance approach-oriented (n=3, $\alpha=0.74$)	3.48 [3.25, 3.71]	2.86 [2.60, 3.13]	<0.0005	0.64
Clerkship learning environment is performance avoid-oriented (n=5, $\alpha=0.86$)	4.18 [4.00, 4.36]	3.80 [3.57, 4.02]	<0.0005	0.49
Student vulnerability to stereotype threat (n=5, $\alpha=0.79$)	2.54 [2.29, 2.79]	2.42 [2.16, 2.69]	0.37	0.12

^aThe authors administered a previously designed survey with six distinct factors relating to perceptions of clerkship grading, as well as adapted mastery, performance approach, and performance avoid Clerkship Goal Structure Scales and Stereotype Vulnerability Scale.^{10,36,37} The authors converted Likert items to a 1 to 5 score. Factor means equaled the average of the items composing the factors.

^bGiven Bonferroni correction for p-value, we report 99.5% confidence intervals instead of 95% CI.

^cFactor means were compared using a non-paired two-tailed t-test with Bonferroni correction. Significant p value <0.005 .

^dCohen’s d was calculated using the pooled standard deviation.

stereotype threat than non UIM students (3.20 vs. 2.13) and perceived more bias in evaluations (3.69 vs. 3.23) while first-generation students experienced more racial/ethnic stereotype threat (3.06 vs. 2.23), compared to non-first generation (p-values all <0.005).

Perceived importance of contributors to final clerkship grades

After-intervention students rated some contributors to grades differently from students with tiered clerkship grades. They rated their improvement over the course of the clerkship (4.89 before vs. 6.74 after, $d=0.75$) and rapport with patients and families (5.43 vs. 6.67, $d=0.48$) as significantly more important in determining their final grades, and written examinations as less important (5.97 vs. 4.92, $d=0.40$) compared to before-intervention students (Table 3). No other contributor had a statistically significant change in perceived importance in determining final grades. Both groups rated being liked by the team, particular residents and attendings worked with, and oral presentations highest in determining their final grades (Table 3).

Recommendations to maximize the fairness of grading under the pass/fail system

Forty-nine students responded to the question on recommendations to improve the pass/fail grading system. Even with this prompt, student responses generally endorsed pass/fail grading and described it as a favorable intervention:

If you showed up every day to learn, put in a good effort, respected patients and your team, and were engaged, and had average-ish fund of knowledge, you

would get the “pass” from attendings/residents. I think this was excellent. Having the pass/fail system allowed me to LEARN rather than worry about performing. (Participant 98)

Some students expressed concerns surrounding evaluator bias with respect to race/ethnicity, gender, LGBTQ identity, or clinical site (32%). Participant recommendations addressed supervisor training, final grades that considered the number of interactions with an assessor and a growth-oriented learning environment that rewards improvement (44%). Because these themes were consistent with those of our previous clerkship grading study, they are not included in Table 4.

In the current study, students expressed heightened concerns about how their performance would be captured in summary narratives (18%) and the MSPE (14%) (Table 4). For instance, students highlighted issues with inconsistent narrative summaries and felt that a site director who knew them personally should author the final narrative. With the MSPE as the primary representation of core clerkship performance, students wanted transparency to ensure that their clerkship data would be conveyed fairly and accurately in a way that allowed them “to stand out in residency applications now that core clerkships are pass/fail.” (Participant 84)

Recommendations to maximize usefulness of feedback

Fifty students provided comments addressing how to maximize the usefulness of feedback on clerkships (Table 4). We grouped comments into those which specifically addressed the WBA intervention and those which recommended how to improve feedback

Table 3. Comparison of students’ perceptions of the determinants of their final clerkship grade from one medical school before and after an assessment and grading intervention.^a

Determinant of Final Clerkship Grade	Before Intervention Mean Rating of Importance (99.5% CI) ^b	After Intervention Mean Rating of Importance (99.5% CI)	p-value ^c	Cohen’s d
Being liked by the team	8.67 [8.16, 9.19]	8.64 [8.20, 9.09]	0.86	0.02
Oral presentations	8.55 [8.11, 8.98]	8.64 [8.31, 8.96]	0.69	0.06
Particular residents that you work with	8.62 [8.13, 9.11]	8.49 [7.96, 9.03]	0.59	0.07
Particular attendings that you work with	8.60 [8.13, 9.06]	8.24 [7.68, 8.80]	0.15	0.18
Working hard	7.44 [6.82, 8.05]	7.77 [7.28, 8.26]	0.25	0.15
Clinical reasoning skills	7.59 [7.11, 8.06]	7.53 [7.09, 7.98]	0.75	0.03
Clinical site at which you do your rotation	7.21 [6.46, 7.97]	7.26 [6.57, 7.96]	0.94	0.02
Fund of knowledge	7.25 [6.80, 7.69]	7.23 [6.84, 7.61]	0.85	0.01
Improvement over the course of the clerkship	4.89 [4.21, 5.57]	6.74 [6.15, 7.33]	<0.0005	0.75
Rapport with patients and families	5.43 [4.67, 6.19]	6.67 [6.11, 7.22]	<0.0005	0.48
Written examination	5.97 [5.33, 6.61]	4.92 [4.19, 5.65]	0.002	0.40

^aOn a scale of zero to ten, students rated how important each contributor was in determining their final grade (0=not important at all and 10=extremely important).

^bBefore-intervention data was previously published as aggregate data as a part of a multi-institutional study on clerkship grading.¹⁰ Given Bonferroni correction for p-value, we report 99.5% confidence intervals instead of 95% CI.

^cThe authors compared means using a non-paired two-tailed t-test with Bonferroni correction with significant p-value of <0.005.

Table 4. Students' recommendations to maximize the fairness of clerkship grading and improve feedback after an assessment intervention that included change to pass/fail grading and increased feedback.^a

Theme	Description of Theme	Representative Quotation	No. (%) of Comments
Recommendations to improve clerkship assessment^b			
Summative evaluation	Students felt unsure of how content was selected for inclusion in the final narrative clerkship evaluation. They desired a more uniform process for writing the summary narratives, and many felt that someone who knew them should author the final evaluation to produce a less generic and more individualized summary narrative.	"The clerkship director writing the summary comments had never even met me and has to write 150 of these over the year. It might result in more personalized, meaningful summary comments if they had each site director write the final summary statement for their own students instead of the overall clerkship director." (Participant 13)	9/49 (18%)
MSPE	Students highlighted a lack of communication and transparency in how narrative comments would be incorporated into the MSPE. There was concern that without grades the final clerkship summaries gathered in the MSPE would not allow students to stand out from peers. Some participants requested inclusion of more direct quotations from supervisor evaluations into the MSPE. A few students noted the importance of regular review of clerkship summary evaluations and MSPEs to ensure equity across race and ethnicity, gender, and LGBTQ status.	"I and many students in my class still feel confused about how the Dean's Letter will [incorporate] our end-evaluations; and how to stand out in residency applications now that core clerkships are pass/fail; and what it means if you happen to get a lack-luster or bad eval from an evaluator, especially if it feels unfair. I think continued communication with students about these issues/questions will help maximize fairness." (Participant 84)	7/49 (14%)
Recommendations to improve clerkship feedback			
Advantages and drawbacks to requiring frequent workplace based assessments (WBAs)	Although students viewed WBAs as an opportunity for formative feedback, weekly WBAs requirements were viewed as administratively burdensome for both students and evaluators, straining the trainee-assessor relationship. Students expressed concerns around evaluator buy-in and evaluation fatigue that led to lower quality evaluations and feedback repetition. Some recommended reducing the number of required WBAs.	"I know [WBAs] are an attempt to make residents/ attendings give you concrete feedback before your final evaluation, but I don't think attendings take it very seriously yet" (Participant 107) "It was awkward, unnecessary, and detracted from the team dynamic, because suddenly I became a clerical burden." (Participant 98)	10/50 (20%)
Opportunities for learner training in feedback process	Students commented that they lacked skills in effectively asking for feedback or seeking clarification of low-quality feedback.	"Not every student knows how to ask for specific feedback, there should be some sort of tool/handout/workshop students can use to help them ask for feedback." (Participant 10)	5/50 (10%)
Opportunities for assessor growth in feedback process	Students felt that delivering feedback should be considered a teachable skill which should be practiced over time. Educator training should include information on bias, appropriate trainee-level expectations, and accountability for poor feedback.	"Allowing students to rate the usefulness of the evaluation ... for their own growth might be helpful, since I've heard of many students [getting] really lazy evaluations from the same people over and over again." (Participant 6)	16/50 (32%)
Characteristics of sub-optimal feedback for student learning	Undesirable feedback cited by students was characterized by lack of in-person delivery of critical feedback, excessive feedback at one time, lack of synthesized feedback on general trajectory, discrepancies among evaluations (oral vs. written, narrative vs. numeric), and seemingly subjective grading criteria.	"Too often evaluations feel like you're being 'back stabbed' by someone who you seemed to have a good working relationship with, only to find out they thought you had a deficiency in knowledge or didn't do something they expected." (Participant 75)	12/50 (24%)
Characteristics of desirable feedback which facilitated growth	Desirable feedback was described as low-stakes, frequent, and rewarding of improvement. Students requested feedback that was based on specific examples with actionable recommendations for improvement. Students consistently advocated for in-person feedback preceding the final written evaluation.	"Substantive feedback with specific examples were always more helpful than vague phrases of compliment... evaluations that cited specific patient interactions or physical exam maneuvers (e.g., should work toward listening for murmurs) [were] more informative than 'need to work on physical exam.'" (Participant 37) "Verbal feedback should begin early in the clerkship, with check ins throughout, so learning goals can be identified early and achievement of goals can be identified." (Participant 117)	31/50 (62%)

^aBased on content analysis of two open-ended questions from a survey of 118 students from one U.S. medical school in 2020.

^bDid not include comments which were consistent with previous multi-institutional study which detailed student recommendations to improve assessment under tiered grading system (see manuscript text for these themes).¹⁰

generally. Some students felt that requirement of two WBAs per week constituted a clerical burden for students or supervisors that interfered with their

relationship with attendings, some of whom did not seem to buy into the new feedback system (20%). Respondents worried that this burden interfered with

feedback quality and final summative evaluations from supervisors:

I started to notice that residents and attendings get to a point of “eval(uation)-fatigue” that undermined the quality of the evaluations, and some may complete the WBA over the final evaluations when pressed for time. (Participant 37)

With respect to feedback, students identified opportunities for both learner and assessor growth. Some students felt they lacked skills to ask for feedback effectively or follow up on low-quality feedback (10%). They also felt that evaluators should be coached on the feedback they provide: “for the feedback that is given—evaluators are not told whether their feedback is helpful or given support in improving it” (Participant 15) (32%). Of the comments which described characteristics of high-quality feedback (62%), students overwhelmingly felt that in-person feedback should precede written feedback, especially for constructive feedback so that students could ask clarifying questions.

Discussion

This before-after survey study found that an assessment system change to pass/fail grading with increased formative feedback was associated with favorable changes in student perceptions of clerkship assessment and seems to have promoted a mastery-oriented clerkship learning environment. Students perceived greater fairness of resident evaluations and overall grading and felt they received more useful feedback compared to students before the intervention. Our previous multi-institutional clerkship grading study showed that small variations in assessment systems (e.g., varied honors caps, normative vs. criterion grading) were not associated with differences in student perceptions.¹⁰ Other studies corroborate the challenges for students navigating variable and seemingly arbitrary expectations for their learning and assessments of their performance.^{43,44} Taken together with the data presented here, our findings suggest that larger assessment system changes using pass/fail grading and WBAs for feedback can address pervasive concerns with core clerkship assessment and grading.¹⁰ Written comments from students revealed improvement but incomplete resolution of tensions around a summative assessment culture with persistent focus on summary narratives and the MPSE.

Under pass/fail grading, students in this study viewed clerkships as more mastery-oriented and less performance approach- and performance avoid-oriented than with tiered grading. After-intervention students scored their own improvement as more important in

determining final grades compared with before-intervention students. This environmental shift toward valuing improvement fosters master adaptive learners who feel empowered to make changes to enhance their performance.⁴⁵ Mastery orientation is associated with self-regulated learning, seeking help when confused, and greater deep processing—behaviors such as questioning knowledge.^{46–48} Our findings support similar results from a qualitative study by Seligman et al, in which interviewed students revealed how removal of tiered grades and increased formative feedback favorably affected their internal, autonomous motivation.²⁸ Reinforcing the pillars of SDT, their autonomous motivation was promoted through enhanced perceptions of autonomy, competence, and relatedness.^{16,28} Our findings reflect increased learner autonomy as students reported that the removal of tiered grades allowed them to focus on their own learning goals as opposed to performing. The measured shift toward mastery learning orientation and improved perceptions of feedback may enhance long-term competence.^{16,49} Of note, increased feedback may include more corrective feedback, which may threaten trainees’ self-perception of competence; even so, the overall shift toward a mastery orientation represents a culture change that can enhance learners’ receptivity to feedback.^{50,51} Because the majority of a physician’s career is spent outside formal education, medical education should facilitate development of lifelong learning practices.⁵² Our findings suggest that our intervention increased autonomous motivation, which would support these practices.

Notably, students’ perceptions of evaluation bias, stereotype threat, and attending evaluation quality did not improve with our intervention. It is noteworthy that UIM and first-generation medical students perceived some aspects of the learning environment less favorably than their peers. Bias and stereotype threat likely reflect persistent underlying medical, social, and institutional power structures, including paucity of residents and attendings from underrepresented backgrounds, insufficient number of women in leadership positions, and unaddressed microaggressions from patients and physicians.^{53–55} The clerkship grading intervention did not explicitly address or rectify these issues. While our institution has developed supervisor diversity, equity, and inclusion trainings to teach cultural humility and mitigate biases, to our knowledge, there are no data confirming that current widely implemented models of bias training (such as incorporating the implicit association test) mitigate assessment biases.^{56,57} Given the continuing perception of bias in evaluation as revealed in this study, more effective interventions are needed to promote equity in assessment.

This assessment system intervention affects the transition from undergraduate to graduate medical education. This transition is currently problematic, with overreliance on grades and standardized tests scores as residency programs struggle to process large numbers of applications.⁵⁸ The previous focus on tiered clerkship grades for before-intervention students shifted for after-intervention students onto summary narratives and the MSPE—the data that would be conveyed to residency programs. Our findings suggest a need for faculty and resident training to write detailed and accurate narrative descriptions of learner performance. Despite national guidelines for the content of the MSPE, variation persists nationally, threatening the utility of these documents for residency programs.⁵⁹ Standardization of the MSPE in a way that also allows for incorporation of additional institution-specific learner performance information may increase its utility, providing richer data than grades alone.⁵⁹

This study has limitations. This was a single-institution survey study and findings may not generalize to other institutional settings. Students self-reported their experiences, which were not confirmed with other measures of performance or opinions of other stakeholders such as supervisors or clerkship directors. This was a before-after study design. Given the high stakes nature of clerkship grades for students, we were not able to use a randomized concurrent groups design. This was a complex system intervention, and it is impossible to fully parse which aspects of the intervention drove student perceptions.

Conclusion

An assessment system intervention involving pass/fail grading and increased formative feedback was associated with moderate to large statistically and practically significant improvements in students' perceptions of the fairness of grading, utility of feedback, and fairness of resident evaluations. After-intervention, students viewed clerkships as more mastery-oriented. The intervention did not improve students' perceptions of evaluation accuracy or bias nor the experience of racial/ethnic stereotype threat. Other medical schools may consider similar interventions to begin to address student concerns with clerkship assessment and promote a more adaptive learning environment.

Acknowledgments

None

Funding

This work was supported by both a UCSF Resident Research Funding Award and by a Mini-Grant from the Association of American Medical Colleges Western Group on Educational Affairs.

Ethical approval

This study was approved by the Institutional Review Board at the University of California, San Francisco (IRB #17-23328).

ORCID

Justin L. Bullock  <http://orcid.org/0000-0003-4240-9798>

Patricia S. O'Sullivan  <http://orcid.org/0000-0002-8706-4095>

Karen E. Hauer  <http://orcid.org/0000-0002-8812-4045>

References

1. Al-Kadri HM, Al-Kadi MT, Van Der Vleuten CPM. Workplace-based assessment and students' approaches to learning: a qualitative inquiry. *Med Teach*. 2013;35(sup1):S31–S38. doi:10.3109/0142159X.2013.765547.
2. Schut S, Driessen E, van Tartwijk J, van der Vleuten C, Heeneman S. Stakes in the eye of the beholder: an international study of learners' perceptions within programmatic assessment. *Med Educ*. 2018;52(6):654–663. doi:10.1111/medu.13532.
3. Bok HG, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Med Educ*. 2013;13(1):123–129. doi:10.1186/1472-6920-13-123.
4. Heeneman S, Oudkerk Pool A, Schuwirth LWT, van der Vleuten CPM, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. *Med Educ*. 2015;49(5):487–498. doi:10.1111/medu.12645.
5. Harrison CJ, Könings KD, Dannefer EF, Schuwirth LWT, Wass V, van der Vleuten CPM. Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. *Perspect Med Educ*. 2016;5(5):276–284. doi:10.1007/s40037-016-0297-x.
6. 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.
7. Ross DA, Boatright D, Nunez-Smith M, Jordan A, Chekroud A, Moore EZ. Differences in words used to describe racial and gender groups in medical student performance evaluations. *PLoS One*. 2017;12(8):e0181659. doi:10.1371/journal.pone.0181659.
8. 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.

9. National Resident Matching Program. *Results of the 2018 NRMP Program Director Survey*. 2018. <https://www.nrmp.org/wp-content/uploads/2018/07/NRMP-2018-Program-Director-Survey-for-WWW.pdf>. Accessed April 28, 2021.
10. 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(11):S48–S56. doi:10.1097/ACM.0000000000002905.
11. 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.
12. Krupat E, Borges NJ, Brower RD, et al. The Educational Climate Inventory. *Acad Med*. 2017;92(12):1757–1764. doi:10.1097/ACM.0000000000001730.
13. Dweck CS. Motivational processes affecting learning. *Am Psychol*. 1986;41(10):1040–1048. doi:10.1037/0003-066X.41.10.1040.
14. Cook DA, Artino AR. Motivation to learn: an overview of contemporary theories. *Med Educ*. 2016;50(10):997–1014. doi:10.1111/medu.13074.
15. Ryan R, Deci E. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol*. 2000;55(1):68–78. doi:10.1037/0003-066X.55.1.68.
16. Ten Cate TJ, Kusurkar RA, Williams GC. How self-determination theory can assist our understanding of the teaching and learning processes in medical education. AMEE guide No. 59. *Med Teach*. 2011;33(12):961–973. doi:10.3109/0142159X.2011.595435.
17. Kusurkar RA. Autonomous motivation in medical education. *Med Teach*. 2019;41(9):1083–1084. doi:10.1080/0142159X.2018.1545087.
18. Kusurkar RA, Ten Cate TJ, Vos CMP, Westers P, Croiset G. How motivation affects academic performance: a structural equation modelling analysis. *Adv Health Sci Educ Theory Pract*. 2013;18(1):57–69. doi:10.1007/s10459-012-9354-3.
19. Kusurkar RA, Croiset G, Galindo-Garré F, Ten Cate O. Motivational profiles of medical students: association with study effort, academic performance and exhaustion. *BMC Med Educ*. 2013;13(1):87. doi:10.1186/1472-6920-13-87.
20. van der Burgt SME, Kusurkar RA, Wilschut JA, Tjin A, Tsoi SLNM, Croiset G, Peerdeman SM. Medical specialists' basic psychological needs, and motivation for work and lifelong learning: a two-step factor score path analysis. *BMC Med Educ*. 2019;19(1):339. doi:10.1186/s12909-019-1754-0.
21. Sockalingam S, Wiljer D, Yufe S, et al. The relationship between academic motivation and lifelong learning during residency: a study of psychiatry residents. *Acad Med*. 2016;91(10):1423–1430. doi:10.1097/ACM.0000000000001256.
22. Hojat M, Nasca TJ, Erdmann JB, Frisby AJ, Veloski JJ, Gonnella JS. An operational measure of physician lifelong learning: its development, components and preliminary psychometric data. *Med Teach*. 2003;25(4):433–437. doi:10.1080/0142159031000137463.
23. Crooks TJ. The impact of classroom evaluation practices on students. *Rev Educ Res*. 1988;58(4):438–481. doi:10.3102/00346543058004438.
24. White CB, Fantone JC. Pass-fail grading: laying the foundation for self-regulated learning. *Adv Health Sci Educ Theory Pract*. 2010;15(4):469–477. doi:10.1007/s10459-009-9211-1.
25. Hauer KE, Lucey CR. Core clerkship grading: the illusion of objectivity. *Acad Med*. 2019;94(4):469–472. doi:10.1097/ACM.0000000000002413.
26. Schneider J. Pass-fail raises the question: what's the point of grades? *New York Times*. <https://www.nytimes.com/2020/06/25/opinion/coronavirus-school-grades.html>. Published June 25, 2020. Accessed June 30, 2020.
27. Harrison CJ, Könings KD, Schuwirth LWT, Wass V, Van der Vleuten CPM. Changing the culture of assessment: the dominance of the summative assessment paradigm. *BMC Med Educ*. 2017;17(1):1–14. doi:10.1186/s12909-017-0912-5.
28. Seligman L, Abdullahi A, Teherani A, Hauer KE. From grading to assessment for learning: a qualitative study of student perceptions surrounding elimination of core clerkship grades and enhanced formative feedback. *Teach Learn Med*. 2020; ePub ahead of print.
29. 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(11):1367–1373.
30. Wasson LT, Cusmano A, Meli L, et al. Association between learning environment interventions and medical student well-being: a systematic review. *JAMA*. 2016;316(21):2237–2252. doi:10.1001/jama.2016.17573.
31. Vokes J, Greenstein A, Carmody E, Gorczyca JT. The current status of medical school clerkship grades in residency applicants. *J Grad Med Educ*. 2020;12(2):145–149. doi:10.4300/JGME-D-19-00468.1.
32. Schopper H, Rosenbaum M, Axelsson 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(1):1–8. doi:10.1186/s12909-016-0813-z.
33. Howley LD, Wilson WG. Direct observation of students during clerkship rotations: a multiyear descriptive study. *Acad Med*. 2004;79(3):276–280.
34. 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.
35. Liaison Committee on Medical Education. Functions and structure of a medical school. <https://lcme.org/publications/>. Accessed June 30, 2020.
36. Artino AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. *Med Teach*. 2014;36(6):463–474. doi:10.3109/0142159X.2014.889814.

37. Midgley C, Maehr ML, Huda LZ. *Manual for the Patterns of Adaptive Learning Scales*. University of Michigan; 2000.
38. Spencer SJ. *The Effect of Stereotype Vulnerability on Women's Math Performance* [doctoral thesis] Ann Arbor, MI: University of Michigan; 1993.
39. Pedhazur EJ, Kerlinger FN. *Multiple Regression in Behavioral Research: Explanation and Prediction*. 2nd ed. New York: Holt, Rinehart, and Winston; 1982.
40. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York, NY: Lawrence Erlbaum Associates; 1988.
41. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15(9):1277–1288. doi:10.1177/1049732305276687.
42. Barry CA, Britten N, Barber N, Bradely C, Stevenson F. Using reflexivity to optimize teamwork in qualitative research. *Qual Health Res*. 1999;9(1):26–44. doi:10.1177/104973299129121677.
43. Han H, Roberts NK, Korte R. Learning in the real place: medical students' learning and socialization in clerkships at one medical school. *Acad Med*. 2015;90(2):231–239. doi:10.1097/ACM.0000000000000544.
44. Larsen DP, Wesevich A, Lichtenfeld J, Artino AR, Brydges R, Varpio L. Tying knots: an activity theory analysis of student learning goals in clinical education. *Med Educ*. 2017;51(7):687–698. doi:10.1111/medu.13295.
45. Cutrer WB, Atkinson HG, Friedman E, et al. Exploring the characteristics and context that allow master adaptive learners to thrive. *Med Teach*. 2018;40(8):791–796. doi:10.1080/0142159X.2018.1484560.
46. Wolters CA, Yu SL, Pintrich PR. The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learn Individ Differ*. 1996;8(3):211–238. doi:10.1016/S1041-6080(96)90015-1.
47. Ryan AM, Pintrich PR. "Should I ask for help?" The role of motivation and attitudes in adolescents' help seeking in math class. *J Educ Psychol*. 1997;89(2):329–341. doi:10.1037/0022-0663.89.2.329.
48. Elliot AJ, McGregor HA, Gable S. Achievement goals, study strategies, and exam performance. *J Educ Psychol*. 1999;91(3):547–563.
49. Dorsey JK, Beason AM, Verhulst SJ. Relationships matter: enhancing trainee development with a (simple) clerkship curriculum reform. *Teach Learn Med*. 2019;31(1):76–86. doi:10.1080/10401334.2018.1479264.
50. ten Cate OTJ. Why receiving feedback collides with self determination. *Adv Health Sci Educ Theory Pract*. 2013;18(4):845–849. doi:10.1007/s10459-012-9401-0.
51. Ramani S, Könings KD, Ginsburg S, van der Vleuten CPM. Meaningful feedback through a sociocultural lens. *Med Teach*. 2019;41(12):1342–1352. doi:10.1080/0142159X.2019.1656804.
52. Teunissen PW, Dornan T. The competent novice: lifelong learning at work. *BMJ*. 2008;336(7645):667–669. doi:10.1136/bmj.39434.601690.AD.
53. Bullock JL, Lockspeiser T, Del Pino-Jones A, Richards R, Teherani A, Hauer KE. They don't see a lot of people my color: a mixed methods study of racial/ethnic stereotype threat among medical students on core clerkships. *Acad Med*. 2020; 95:S58–S66. doi:10.1097/ACM.0000000000003628.
54. Larson AR, Kan CK, Silver JK. Representation of women physician deans in U.S. medical schools. *J Women's Heal*. 2019;28(5):600–605. doi:10.1089/jwh.2018.7448.
55. Wheeler M, de Bourmont S, Paul-Emile K, et al. Physician and trainee experiences with patient bias. *JAMA Intern Med*. 2019;179(12):1678–1685. doi:10.1001/jamainternmed.2019.4122.
56. Diaz T, Navarro JR, Chen EH. An institutional approach to fostering inclusion and addressing racial bias: implications for diversity in academic medicine. *Teach Learn Med*. 2020;32(1):110–116. doi:10.1080/10401334.2019.1670665.
57. Sukhera J, Wodzinski M, Rehman M, Gonzalez CM. The implicit association test in health professions education: a meta-narrative review. *Perspect Med Educ*. 2019;8(5):267–275. doi:10.1007/s40037-019-00533-8.
58. Weissbart SJ, Kim SE, Feinn RS, Stock JA. Relationship between the number of residency applications and the yearly match rate: time to start thinking about an application limit? *J Grad Med Educ*. 2015;7(1):81–85. doi:10.4300/JGME-D-14-00270.1.
59. Hauer KE, Giang D, Kapp ME, Sterling R. Standardization in the MSPE: Key Tensions for Learners, Schools, and Residency Programs. *Acad Med*. 2021;96(1):44–49.