Clerkship Grading Committees: the Impact of Group Decision-Making for Clerkship Grading

Annabel K. Frank, MD[1,2], Patricia O'Sullivan, EdD[1], Lynnea M. Mills, MD[1], Virginie Muller-Juge, MSc[1], and Karen E. Hauer, MD, PhD[1]

[1]Department of Medicine, University of California, San Francisco, San Francisco, CA, USA; [2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

BACKGROUND: Faculty and students debate the fairness and accuracy of medical student clerkship grades. Group decision-making is a potential strategy to improve grading.

OBJECTIVE: To explore how one school's grading committee members integrate assessment data to inform grade decisions and to identify the committees' benefits and challenges.

DESIGN: This qualitative study used semi-structured interviews with grading committee chairs and members conducted between November 2017 and March 2018.

PARTICIPANTS: Participants included the eight core clerkship directors, who chaired their grading committees. We randomly selected other committee members to invite, for a maximum of three interviews per clerkship.

APPROACH: Interviews were recorded, transcribed, and analyzed using inductive content analysis.

KEY RESULTS: We interviewed 17 committee members. Within and across specialties, committee members had distinct approaches to prioritizing and synthesizing assessment data. Participants expressed concerns about the quality of assessments, necessitating careful scrutiny of language, assessor identity, and other contextual factors. Committee members were concerned about how unconscious bias might impact assessors, but they felt minimally impacted at the committee level. When committee members knew students personally, they felt tension about how to use the information appropriately. Participants described high agreement within their committees; debate was more common when site directors reviewed students' files from other sites prior to meeting. Participants reported multiple committee benefits including faculty development and fulfillment, as well as improved grading consistency, fairness, and transparency. Groupthink and a passive approach to bias emerged as the two main threats to optimal group decision-making.

CONCLUSIONS: Grading committee members view their practices as advantageous over individual grading, but they feel limited in their ability to address grading fairness and accuracy. Recommendations and support may help committees broaden their scope to address these aspirations.

KEY WORDS: medical education-qualitative methods; medical education-undergraduate; evaluation; clerkship grading; group decision-making; grading committees; clinical competence.

J Gen Intern Med 34(5):669–76
DOI: 10.1007/s11606-019-04879-x

We presented an earlier version of the manuscript as a mini-oral presentation at the UCSF Education Showcase in San Francisco, CA, in May 2018.

Electronic supplementary material *The online version of this article (https://doi.org/10.1007/s11606-019-04879-x) contains supplementary material, which is available to authorized users.*

Published online April 16, 2019

INTRODUCTION

Assigning medical student clerkship grades is a critical task for educators to demonstrate to students the quality of their learning and performance. Program directors rely on clerkship grades during residency selection.[1, 2] However, students and faculty alike question the fairness and accuracy of clerkship grades.[3–6] Attempts to optimize grade assignments through standard policies, tools, or rater training have thus far been imperfect.[7–9] Group decision-making is a potential strategy to improve the process of interpreting evaluation information from multiple supervisors and assigning clerkship grades.[10]

Clinical performance assessment entails collecting multiple sources of information that experts review to make competence judgments.[11, 12] These judgments entail generalizing from assessment data, both numerical and narrative, to extrapolate to other contexts.[13, 14] Workplace-based assessments from clinical supervisors, including ratings and comments from attendings and residents, usually constitute the major data for clerkship grades.[13] As with any judgment of human performance, achieving acceptable inter-rater reliability is challenging even with training.[7, 8, 15–17] Variability arises because assessors notice and value different aspects of performance to different degrees and form unique global impressions.[16] Impressions of learner performance are influenced by unconscious biases, and differences in clerkship grades reported by students who are non-white, male, or reticent heighten concern about fairness.[18–24] These issues illustrate the complexity of information that must be considered when assigning clerkship grades.

Graduate medical education requires group decision-making within clinical competency committees (CCCs) for review of resident performance.[25] The rationale for CCCs stems from findings that groups drawing on members' knowledge and

experience through information sharing procedures make better decisions than individuals alone.[26–29] However, rigid individual preferences or high desires for conformity can jeopardize the quality of group decisions.[30] Early descriptions of CCCs highlight how members contemplate which data to use and weigh usefulness of that data.[31–33] Although clerkship committees differ from CCCs in terms of their data, learner characteristics, and purpose, similar lessons and challenges could emerge.

From a theoretical perspective, the social decision scheme (SDS) theory describes how individuals share information to reach a decision; key elements are individual preferences about the decision, group composition (distribution of preferences within the group), group influences (procedures to reach a decision), and collective response (the decision).[34] According to SDS theory, a powerful influence on decisions is the group's starting point, characterized by initial individual preferences and group composition. Constructivist theory explicates how a group comes to shared understanding of a learner's performance.[34, 35] Recommendations for optimal committee functioning include selecting heterogeneous members, minimizing time pressure, and fostering an environment and implementing procedures that promote sharing information and perspectives.[10]

Although group evaluation sessions have been described for evaluating students' progress and judging marginal student performance in internal medicine clerkships, we are unaware of any description of grading committee processes across disciplines.[36, 37] This study aims to: (1) explore how clerkship grading committees use assessment data to make group judgments about student performance and assign grades; and (2) identify committee members' perspectives of grading committee benefits and challenges. Findings will inform educators about the process, potential benefits, and challenges of clerkship grading committees.

METHODS

Study Design

This is a qualitative study using a conventional, inductive approach to identify themes and perspectives expressed by participants.[38] The University of California, San Francisco (UCSF), Institutional Review Board approved this study as exempt.

Setting

In 2016, UCSF introduced a requirement for grading committees in all eight core clerkships (anesthesia, family and community medicine, internal medicine, neurology, obstetrics-gynecology, pediatrics, psychiatry, surgery). One clerkship (internal medicine) had used a committee since 2008, and in other clerkships, the site (SD) or clerkship (CD) directors had assigned grades. Clerkships span 2 to 8 weeks. Requirements standardized across clerkships were committees synthesize assessments completed by attendings and residents and assign a clerkship grade for each student (honors/pass/fail). The assessment form includes 10 competency-based items scored 1–4, narrative comments fields, a reporter/interpreter/manager/educator (RIME) rating, and a confidential grade recommendation (Appendix 1, online). Each committee comprises three or more faculty members familiar with student education, including the CD. Committees were expected to convene after all assessments were submitted but prior to 6 weeks after the clerkship and use recommended evidence-based guidelines (Appendix 2, online). Procedures and data synthesis could vary by clerkship.

Sample

We invited all eight CDs, who chair their respective grading committees, to participate in individual semi-structured interviews. We also invited two other randomly selected committee members from each clerkship. If a member declined, we randomly selected another committee member.

Data Collection

The research team (AKF, PO'S, KEH) developed the interview guide based on literature about clerkship grading, CCCs, and group decision-making (Appendix 3, online).[3, 10–13, 16–18, 31] Questions explored committee procedures and decision-making processes, member training, committee pre-meeting work, typical and difficult grading discussions, perceived impact of bias on grading, and committee benefits and challenges. The interviewer used follow-up questions and probes to explore fully each topic. Demographic questions addressed the role in clerkship, faculty rank, gender, race, and ethnicity.

Based on two pilot interviews with faculty who were not committee members but had experience in grading and assessing students, we edited the guide for clarity. Potential participants received an email invitation. We emailed a consent document in advance, discussed it before interviews, and obtained verbal consent. Interviews occurred between November 2017 and March 2018. One trained researcher (AKF), a fourth-year student from a different medical school, conducted all interviews that lasted 42–87 min and occurred in person (2) or by phone (15). Interviews were recorded, professionally transcribed, and de-identified before analysis. Participants received no compensation.

Data Analysis

We calculated descriptive statistics for participants' characteristics. Analysis occurred concurrently with data collection, and we deemed 17 interviews sufficient for saturation.[39] Two authors (AKF, KEH) closely read the first two transcripts, identifying concepts to inform codebook development. Three authors (AKF, LMM, KEH) read three more transcripts and refined initial concepts into a codebook. Two investigators (AKF, LMM) coded each interview independently and reconciled differences through discussion, or as needed with a third investigator (KEH). All five team members read and summarized excerpts for each code to identify larger themes within and across codes. We used constant comparison to refine

properties of each theme.[40] Sensitizing concepts from literature on group decision-making and clerkship grading guided our analysis.[38, 41] We used Dedoose software (Los Angeles, CA, USA, Version 8.0.36) for coding, organizing, and retrieving data. Throughout the study, we engaged in reflexivity through journaling and discussions to maintain awareness of our own reactions and potential biases.[42]

RESULTS

Participants

Twenty-five faculty grading committee members received email invitations. Two did not respond, four declined, and two accepted but failed to schedule interviews. Seventeen (68%) participated (Table 1). Participants included all eight CDs.

Results below describe committees: characteristics, decision-making procedures, and data sources. We then discuss five themes: information from individual assessors, valuing competencies, bias, resolving disagreements, and committee impact. For confidentiality, we assigned committees letters (A–H) and individual numbers to participants.

Committees

Committee Characteristics. Most committees met once per clerkship block for 30–90 min with 3–11 members (Appendix 4, online). Many committees included only the CD and SDs, and some also included administrators or other faculty interested in education. In clerkship D, the CD and assistant CD alone determined grades.

Table 1 Participant Characteristics ($N = 17$)

Characteristic	No. of subjects (%)*
Specialty	
Anesthesia	1 (6)
Family medicine	2 (12)
Internal medicine	3 (18)
Neurology	3 (18)
Obstetrics and gynecology	1 (6)
Pediatrics	2 (12)
Psychiatry	3 (18)
Surgery	2 (12)
Gender	
Female	10 (59)
Academic rank	
Professor	7 (41)
Associate professor	4 (24)
Assistant professor	6 (35)
Role in clerkship	
Clerkship director	8 (47)
Assistant clerkship director	1 (6)
Site director	8 (47)
Race	
Asian	2 (12)
Black	1 (6)
White	12 (70)
Mixed race	2 (12)
Ethnicity	
Non-Hispanic	17 (100)

**Percentages do not add to 100% due to rounding*

Decision-making Procedures. Committees varied in how they identified students for discussion. Before meeting, most SDs reviewed students from their own sites and/or selected students from other sites and determined preliminary grade recommendations. During the committee meeting, members discussed students for whom there was discrepancy or uncertainty in the grade. Because of an institutional standard for the maximum number of honors, one committee convened every few blocks to discuss retrospectively students who had received pass grades but seemed eligible for honors. All committees focused deliberations on students at the border of receiving honors.

Committees rarely discussed failing grades, largely because few students failed. Some participants explained that they identified struggling students through direct feedback from the student's team to the CD or SD before scheduled committee meetings, and thus convened ad hoc committees to make those grade decisions.

Data Sources. Committees prioritized available data differently. Within and across committees, participants variably considered assessor's narrative comments, numerical scores, RIME ratings, exam, and honors recommendations to make decisions. Despite departmental numerical criteria/guidelines for honors, committees viewed grade decisions as requiring judgment, particularly for students close to honors. Typically, committees started with one data source and then used other sources to corroborate it. One participant reflected on their "non-scientific" data synthesis: "one person may believe in the specificity of numbers, another person may believe in the value of RIME, another person believes in the value of adjectives and how strongly people use them" (B348, SD). Considering assessor identity, time spent with student, and context, they inferred a judgment about the student's true performance.

Narrative Comments. On one committee, all three participants viewed narrative comments as the primary driver for decisions. For two other committees, narrative comments were the only data reviewed during meetings. Comments most readily directed grade decisions when they contained superlative language and/or described specific behaviors. Participants felt comfortable identifying honors students when all assessors "gushed" about a student (A724, SD). The absence of high-praise, or "lukewarm" language (C987, SD), could be the only indication of poor performance: "you kind of read into the absence of glowing remarks as being a negative comment. But maybe that evaluator was just having a bad day" (A724, SD).

Several strategies enabled participants to elucidate whether a terse comment had a "veiled meaning" (A698, SD). Participants affirmed that, with experience, they learned assessors' use of language. Some committees relied on SDs to interpret comments written by assessors at their site. Participants expressed concern about interpreting comments: "by definition, you can't have a shared mental model unless people talk

about the model. We do have it in writing, but we will still debate it when we see the data. What one person finds exceptional reasoning isn't always the same as another person" (B348, SD).

Numerical Scores. Participants from two clerkships cited numerical scores as the primary data source driving decision-making and used narrative comments to confirm impressions. One participant explained, "a lot of this is decided upfront numerically. I don't start from the comments and just say, 'gee, how are these comments?'" (D359, CD). Rather than a strict numerical cut-off, participants analyzed all aspects of the assessment to contextualize the scores. Yet, some participants questioned the accuracy of scores and found them challenging to interpret, because they were not confident all raters observed the students using the skills or knew how to rate them. Another participant expressed concern that "hawks" (supervisors who rate students lower) were the only ones filling out forms accurately (G815, CD).

RIME Ratings. Participants almost unanimously described limitations of RIME ratings. They perceived that the problem was not the RIME terminology but that assessors used it inappropriately: "if they thought it was a good student, they put an E. Which is kind of ridiculous" (G815, CD). Some described it as "useless" (A698, SD) and "totally random" (A724, SD). Despite these challenges, all committees still reviewed RIME adjectives, usually to identify general patterns within an assessment.

Exam. Across committees, participants described using exam scores to corroborate information from assessors, e.g., to confirm hints in narrative comments and scores about below-average medical knowledge.

Themes

Information from Individual Assessors. The first theme addressed varied weights assigned to individual assessors. Several committees numerically weighted attending scores greater than residents'. Committee members considered how to value attending and resident narrative comments differently. Within and across committees, they endorsed conflicting opinions of the relative merit of resident assessments. They universally valued that residents directly observe students interact with patients and think through problems, whereas attendings may only observe students present cases on rounds. Some participants, however, cautioned that residents could be swayed by students who were hardworking, enthusiastic, and personable. When making comparisons to prior students, attendings' comments outweighed residents' comments based on attendings' experience.

Valuing Competencies. The second theme addressed the complexity of considering various competencies. All participants expressed that concerning professionalism (lateness, disappearing) and engagement (disrespectful, lack of enthusiasm) signified red flags that rendered students ineligible for honors. Other competencies were valued differently across and sometimes within clerkships. Most participants from non-procedural specialty committees were reluctant to award honors to students with any signs that they were not outstanding in their medical knowledge and clinical reasoning. However, all three participants from one non-procedural committee expressed unique preferences: one said no competency was more important than another, one was most impressed by humanism, and one thought medical knowledge and patient care distinguished honors from pass students. In contrast, two participants from a procedural specialty were most impressed by work ethic, assertiveness, and confidence—especially when the student was confident and correct.

Bias. All participants felt that unconscious bias could affect assessments but were uncertain how to approach the problem. Participants felt minimally vulnerable to bias at the committee level because they usually did not interact with the students clinically and did not interact with certain students at all. One participant contemplated: "how would I know if a student were under-represented in medicine?" (D359, CD). Some participants or whole committees removed student names when reviewing assessments to help avoid bias based on gender or race/ethnicity. Conversely, others used their personal knowledge to try to mitigate biased decision-making. However, they felt conflicted about whether using knowledge about a student's background was appropriate: "one of our SDs is more apt to bring up someone's…ethnic background or their home background, saying, 'Well, this could account for what we found on a spreadsheet or comments.' Whereas other SDs will say, 'We don't want to let that information enter into our grading'" (B348, SD).

Participants reflected that possible bias related to gender arose occasionally in meetings. Narrative comments such as, "this female student is surprisingly more proactive" might cause them to "hit pause" or exclude an evaluation (E782, CD). One obstetrics-gynecology participant noticed less opportunity for male students to participate with patients and demonstrate skills. Some other participants worried about possible gender bias, although one reflected that bias could be directed against males or females "because it may vary on the team structure" (C710, CD).

Committee members wholeheartedly endorsed a risk of bias against students with certain personality traits in clerkship grading. They perceived extroverted "go-getters" (H512, CD) as having an advantage, while quieter students potentially suffered lower grades. Committee members felt powerless after the fact if students had not impressed their assessors

during the rotation: "fluent, socially aware, comfortable-presenting—those students end up demonstrating more competency. I don't know how to control for that" (G815, CD).

Resolving Disagreements. Committees could almost always avoid disagreements or readily resolve them. Committee members reported high agreement within their committees and confidence in their ability to make consistent decisions. When asked, most participants endorsed having built a strong, shared mental model within their committees of honors versus pass performance; nonetheless, they acknowledged their models' limitations. The most longstanding committee employed multiple recommended practices (Appendix 5, online) including pre-review and formally logging previous difficult decisions as precedent for future decisions. Two committees in which members independently reviewed students in advance of meetings reported the most debate during meetings, prompted by discrepancies in preliminary grade assignments.

Absent consensus, participants across committees reported comfort deferring to either the majority or the SD, or occasionally the CD. Participants trusted SDs to know their site's assessors, team dynamics, and other factors that influenced the student's interactions. One member routinely deferred to the SD: "I basically remain fairly silent because I have no idea who these other students are" (H776, SD). Committee indecision sometimes prompted members to contact assessors for additional information to resolve discrepant numerical scores and narrative comments, inconsistencies, missing assessment data, or scant, vague comments. Participants described grade deadlines as a major constraint to gathering more information.

Committee Impact. This theme represents an endorsement of committees because of perceived transparency and strength of multiple perspectives despite data limitations. Participants identified valuable committee practices (Appendix 5, online); nevertheless, they expressed uncertainty about whether they were rendering the right decisions. Several questioned whether honors grades predicted students' future success better than pass grades. Though many believed their committee had a shared understanding of an honors student, they acknowledged a degree of arbitrariness: "little difference in comments and your evaluations are going to make the difference when you're in that tight narrow range" (F955, CD). Despite their interest in using criterion-based grading, they often compared students to one another. Committee members felt that the strength of the incoming data limited decision accuracy: "you can only argue the merits of imperfect information for so long" (B348, SD). Nevertheless, they believed committees improve data interpretation: "multiple people thinking about each student is more likely to get you an accurate representation" (G828, SD). Participants agreed committees promoted grading consistency: they "made the process more consistent across sites" (B553, CD). Participants believed the committees offered greater transparency to students, especially those who appealed or questioned grades. Members could explain to students that the committee uses a consistent, thoughtful process to establish each decision. Committee decision-making relieved CDs and SDs of personal accountability burden, providing greater peace of mind.

Some participants felt the grading committee operated relatively efficiently, though likely less efficiently than assigning grades individually. Challenges included gathering enough information on time and scheduling committee meetings. However, most felt the committee was worth the extra effort because it fostered faculty development and fulfillment: "grading committee is probably one of these rare instances where more work has been created, and I'm so glad that it has, because it makes me feel better about the product we create and it builds community with the people I work with" (B348, SD).

DISCUSSION

This study reveals that grading committees appear to support members' accountability for their grade assignments, and greater discussion among members to produce a more standardized process for assigning clerkship grades that provides assurance to their committee members. However, despite common guidelines, committee procedures vary across clerkships. Study participants endorsed that committees promote consistency within departments, shared perspectives, and improved transparency for students, with multiple recommended best practices used variably (Appendix 5, online). However, the themes highlight challenges related to prioritizing different data sources and competencies and minimizing bias.

The greatest committee challenge was synthesizing discrepant, and often cryptic, incoming data. Rater variability is inherent to observation and occurs by multiple mechanisms: attributing different importance to each aspect of performance, comparing to personal versions of an ideal student, and synthesizing thoughts to form a distinctive impression.[16] In parallel, this variability has been demonstrated at the committee level, when committee members integrate and judge complex data.[11] Participants expressed desire to eliminate variability in incoming data, and among themselves; ironically, cultivating a heterogeneous group with unique values and perspectives actually improves group performance.[10] They struggled with how much to incorporate personal knowledge of students or students' teams, because that information seemed helpful but was not always available to the committee. These challenges demonstrate that although committees were charged with and aspired to a shared mental model of an honors or pass student, variability in the available data and members' interpretation of that data persisted.

Part of the synthesis challenge rested in considering the data. Although participants had a general approach, their methods were not consistent. They valued data sources to differing degrees, depending on from whom data originated and whether data fit with an emerging global impression. This approach may

introduce confirmation bias, which entails ignoring or justifying data that conflict with one's initial impression. The non-algorithmic approach arises out of concern about the quality of assessments, which is not unfounded. Assessors do not give thought to each component and domain of assessment, but rather convey their impression of a student's overall performance or likability.[43, 44] Committee members felt consistent and experienced in decoding comments, an impression that the literature reports.[45] However, comments that are not constructive compel committees to interpret the underlying truth, and faculty may engage in assessment practices that they lament in other faculty.[45] For example, despite participants' confidence in their decoding skills, our findings suggest they are liable to interpret data to fit their early impressions or information they trust, without thoroughly considering all possibilities. Despite wanting to rely on data objectively, committees and individual members develop their own guidance about which assessment data are most valid, which assessors are most insightful, reasons for inconsistency across data, and the influence of contextual circumstances. Though participants believed the committees improved fairness and transparency, our findings suggest that students understandably may be uncertain about which data contribute to grades and how. Students may find it troublesome that so many different interpretations of data are possible and question the accuracy of data integration.

The committee design reveals heavy reliance on trust of key personnel such as SDs. On some committees, SDs make decisions mostly independently using knowledge about the assessor, student, and context. The group trusts each SD's rationale and expertise, and members may not contribute divergent perspectives. Similarly, CCC faculty are uncomfortable making decisions without personal knowledge of trainees, even though they recognize incoming data as more important.[31]

Participants described openness and respect for different opinions as their committees' strengths; however, descriptions of typical discussions revealed they valued high levels of group agreement to the point of vulnerability. Group harmony is a hallmark of "groupthink," which results when members suppress critical or dissenting thoughts unintentionally.[46] As groups become more cohesive, this conformity heightens.[46] The danger is not that members will fail to object to what others propose, but that they inherently trust other members' decisions without scrutiny. In our study, groupthink emerged as a threat to optimal committee functioning, although committees that review all students in advance appear less vulnerable. Through individual review, alternative interpretations of data arise organically, which allows more meeting time for discussion of the appropriateness of assumptions and subtle data overlooked. This strategy is aligned with recommendations that committees maximize time spent sharing information and perspectives.[10, 47]

A main benefit of the committee system is addressing potential biases in grading, and committee members somewhat saw this as their duty. Lack of direct contact with students led participants to feel shielded against bias, but literature across disciplines suggests significant bias can nonetheless occur.[48] Studies demonstrating that underrepresented minority students, Asians, men, and introverted students earn lower grades in certain clerkships oblige us to take an active approach to understanding where bias occurs and how to counteract it.[18, 23] For example, committee members can recognize that favoring traditional metrics (i.e., exam scores) and medical knowledge could prompt committees to undervalue unique contributions of underrepresented minority students, e.g., related to communication and care coordination with patients from marginalized populations.[24, 49] Though committee members believe exam scores minimally influence their decisions, they may underestimate this influence, which could perpetuate bias.[50] Committees, like CCCs, feel tension about how to prevent bias: whether to blind to assessor and student characteristics or use their personal knowledge of learners.[31] Committees may benefit from routinely discussing their own potential biases and preferences.

Our study has limitations. Findings from this single-institution study with primarily white participants may not be generalizable. We did not interview all grading committee members. We did not examine students' data to attempt to determine the correctness of grade assignments, or follow their performance longitudinally to determine long-term performance outcomes. We did not interview students to corroborate committee members' impressions of committee benefits such as transparency.

These findings demonstrate that grading committees likely promote accountability of committee members, shared understanding of their charge, and consistency within committees, but improving fairness is an ongoing challenge. Multiple recommendations can strengthen committee procedures (Appendix 5, online). Committees may benefit from taking a more active approach to mitigate bias related to student characteristics and tensions regarding the importance of personal knowledge of students. Further research through observations of meetings before and after interventions to apply recommended strategies for group decision-making could clarify benefits of proposed best practices. Given the importance of clerkship grades for students' residency match and the substantial faculty effort to assign grades, committees need results from further studies to optimize the fairness, accuracy, and efficiency of clerkship grading.

Corresponding Author: *Karen E. Hauer, MD, PhD; Department of Medicine University of California, San Francisco, San Francisco, CA, USA (e-mail: karen.hauer@ucsf.edu).*

Funding Information *All funding for this project was provided by the University of California, San Francisco, School of Medicine.*

Compliance with Ethical Standards:

The University of California, San Francisco (UCSF), Institutional Review Board approved this study as exempt. We emailed a consent

document in advance, discussed it before interviews, and obtained verbal consent.

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Publisher's Note: *Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.*

REFERENCES

1. National Resident Matching Program. Data Release and Research Committee. Results of the 2016 NRMP Program Director Survey. http://www.nrmp.org/wp-content/uploads/2016/09/NRMP-2016-Program-Director-Survey.pdf. Accessed December 8, 2018.
2. **Cullen MW, Reed DA, Halvorsen AJ, et al.** Selection criteria for internal medicine residency applicants and professionalism ratings during internship. Mayo Clin Proc. 2011;86(3):197–202.
3. **Alexander EK, Osman NY, Walling JL, Mitchell VG.** Variation and Imprecision of Clerkship Grading in U.S. Medical Schools. Acad Med. 2012;87(8):1070–6.
4. **Kogan JR, Conforti LN, Iobst WF, Holmboe ES.** Reconceptualizing Variable Rater Assessments as Both an Educational and Clinical Care Problem. Acad Med. 2014;89(5):721–7.
5. **Goldstein SD, Lindeman B, Colbert-Getz J, et al.** Faculty and resident evaluations of medical students on a surgery clerkship correlate poorly with standardized exam scores. Am J Surg. 2014;207(2):231–5.
6. **Takayama H, Grinsell R, Brock D, Foy H, Pellegrini C, Horvath K.** Is it Appropriate to Use Core Clerkship Grades in the Selection of Residents? Curr Surg. 2006;63(6):391–6.
7. **Zaidi NLB, Kreiter CD, Castaneda PR, et al.** Generalizability of Competency Assessment Scores Across and Within Clerkships: How Students, Assessors, and Clerkships Matter. Acad Med. 2018;93(8)1212–7.
8. **Pelgrim EAM, Kramer AWM, Mokkink HGA, van den Elsen L, Grol RPTM, van der Vleuten CPM.** In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ. 2011;16(1):131–42.
9. **Kogan JR, Holmboe ES, Hauer KE.** Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 2009;302(12):1316–26.
10. **Hauer KE, Ten Cate O, Boscardin CK, et al.** Ensuring Resident Competence: A Narrative Review of the Literature on Group Decision Making to Inform the Work of Clinical Competency Committees. J Grad Med Educ. 2016;8(2):156–64.
11. **Oudkerk Pool A, Govaerts MJB, Jaarsma DADC, Driessen EW.** From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. Adv Health Sci Educ. 2018;23(2):275–87.
12. **Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA.** Using In-Training Evaluation Report (ITER) Qualitative Comments to Assess Medical Students and Residents: A Systematic Review. Acad Med. 2017;92(6):868–79.
13. **Hemmer PA, Papp KK, Mechaber AJ, Durning SJ.** Evaluation, Grading, and Use of the RIME Vocabulary on Internal Medicine Clerkships: Results of a National Survey and Comparison to Other Clinical Clerkships. Teach Learn Med. 2008;20(2):118–26.
14. **Wass V, Van der Vleuten C, Shatzer J, Jones R.** Assessment of clinical competence. The Lancet. 2001;357(9260):945–49.
15. **Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ.** Constructing a Validity Argument for the Mini-Clinical Evaluation Exercise: A Review of the Research: Acad Med. 2010;85(9):1453–61.
16. **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. 2013;18(3):325–41.
17. **Carline JD, Paauw DS, Thiede KW, Ramsey PG.** Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med. 1992;7(5):506–10.
18. **Lee KB, Jeffe DB.** "Making the Grade:" Noncognitive Predictors of Medical Students' Clinical Clerkship Grades. J Natl Med Assoc. 2007;99(10):1138–50.
19. **Noureddine L, Medina J.** Learning to Break the Shell: Introverted Medical Students Transitioning into Clinical Rotations. Acad Med. 2018;93(6);822
20. **Schuh LA, London Z, Neel R, et al.** Education Research: Bias and poor interrater reliability in evaluating the neurology clinical skills examination. Neurology. 2009;73(11):904–8.
21. **Riese A, Rappaport L, Alverson B, Park S, Rockney RM.** Clinical Performance Evaluations of Third-Year Medical Students and Association With Student and Evaluator Gender: Acad Med. 2017;92(6):835–40.
22. **Lee V, Brain K, Martin J.** Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications: A Systematic Literature Review. Acad Med. 2017;92(6):880–7.
23. **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(5):659–65.
24. **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.
25. Accreditation Council for Graduate Medical Education. Common Program Requirements. http://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf. Accessed December 8, 2018.
26. **Surowiecki J.** The Wisdom of Crowds. Knopf Doubleday Publishing Group; 2005.
27. **Michaelsen LK, Watson WE, Black RH.** A Realistic Test of Individual Versus Group Consensus Decision Making. J Appl Psychol. 1989;74(5):834–9.
28. **Kerr NL, Tindale RS.** Group Performance and Decision Making. Annu Rev Psychol. 2004;55(1):623–55
29. **Klocke, U.** How to improve decision making in small groups: Effects of dissent and training interventions. Small Group Res. 2007;38(3):437–68.
30. **Beran TN, Kaba A, Caird J, McLaughlin K.** The good and bad of group conformity: a call for a new programme of research in medical education. Med Educ. 2014;48(9):851–9.
31. **Ekpenyong A, Baker E, Harris I, et al.** How do clinical competency committees use different sources of data to assess residents' performance on the internal medicine milestones? A mixed methods pilot study. Med Teach. 2017;39(10):1074–83.
32. **Donato AA, Alweis R, Wenderoth S.** Design of a clinical competency committee to maximize formative feedback. J Community Hosp Intern Med Perspect. 2016;6(6):33533.
33. **Schumacher DJ, King B, Barnes MM, et al.** Influence of Clinical Competency Committee Review Process on Summative Resident Assessment Decisions. J Grad Med Educ. 2018;10(4):429–37.
34. **Stasser G.** A Primer of Social Decision Scheme Theory: Models of Group Influence, Competitive Model-Testing, and Prospective Modeling. Organ Behav Hum Decis Process. 1999;80(1):3–20.
35. **Chahine S, Cristancho S, Padgett J, Lingard L.** How do small groups make decisions? Perspect Med Educ. 2017;6(3):192–8.
36. **Gaglione MM, Moores L, Pangaro L, Hemmer PA.** Does Group Discussion of Student Clerkship Performance at an Education Committee Affect an Individual Committee Member's Decisions? Acad Med. 2005;80(10 Suppl):S55–8.
37. **Battistone MJ, Milne C, Sande MA, Pangaro LN, Hemmer PA, Shomaker TS.** The Feasibility and Acceptability of Implementing Formal Evaluation Sessions and Using Descriptive Vocabulary to Assess Student Performance on a Clinical Clerkship. Teach Learn Med. 2002;14(1):5–10.
38. **Hsieh H-F, Shannon SE.** Three Approaches to Qualitative Content Analysis. Qual Health Res. 2005;15(9):1277–88.
39. **Morse JM.** The significance of saturation. Qual Health Res. 1995;5(2);147–9.
40. **Glaser BG, Strauss AL.** The constant comparative method in qualitative analysis. In: The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago: Aldine Transaction; 1967.
41. **Bowen GA.** Grounded Theory and Sensitizing Concepts. Int J Qual Methods. 2006;5(3):12–23.
42. **Barry CA, Britten N, Barber N, Bradley C, Stevenson F.** Using Reflexivity to Optimize Teamwork in Qualitative Research. Qual Health Res. 1999;9(1):26–44.
43. **Tavares W, Ginsburg S, Eva KW.** Selecting and Simplifying: Rater Performance and Behavior When Considering Multiple Competencies. Teach Learn Med. 2016;28(1):41–51.
44. **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(4):335–41.
45. **Ginsburg S, Regehr G, Lingard L, Eva KW.** Reading between the lines: faculty interpretations of narrative evaluation comments. Med Educ. 2015;49(3):296–306.

46. **Janis IL.** Groupthink. Psychol Today 1971;5:43–6, 74-6.
47. **Kinnear B, Warm EJ, Hauer KE.** Twelve tips to maximize the value of a clinical competency committee in postgraduate medical education. Med Teach. 2018.
48. **Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J.** Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci. 2012;109(41):16474–9.
49. **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(11):1472–4.
50. **Lurie SJ, Mooney CJ.** Assessing a Method to Limit Influence of Standardized Tests on Clerkship Grades. Teach Learn Med. 2012;24(4):287–91.