Multi-institutional Implementation of the National Clinical Assessment Tool in Emergency Medicine: Data From the First Year of Use

Katherine Hiller, MD, MPH[1], Julianna Jung, MD[2], Luan Lawson, MD[3], Rebecca Riddell, MS[4], and Doug Franzen, MD, MEd[5]

ABSTRACT

Objectives: Uniformly training physicians to provide safe, high-quality care requires reliable assessment tools to ensure learner competency. The consensus-derived National Clinical Assessment Tool in Emergency Medicine (NCAT-EM) has been adopted by clerkships across the country. Analysis of large-scale deidentified data from a consortium of users is reported.

Methods: Thirteen sites entered data into a Web-based platform resulting in over 6,400 discrete NCAT-EM assessments from 748 students and 704 assessors. Reliability, internal consistency analysis, and factorial analysis of variance for hypothesis generation were performed.

Results: All categories on the NCAT-EM rating scales and professionalism subdomains were used. Clinical rating scale and global assessment scores were positively skewed, similar to other assessments commonly used in emergency medicine (EM). Professionalism lapses were noted in <1% of assessments. Cronbach's alpha was >0.8 for each site; however, interinstitutional variability was significant. M4 students scored higher than M3 students, and EM-bound students scored higher than non–EM-bound students. There were site-specific differences based on number of prior EM rotations, but no overall association. There were differences in scores based on assessor faculty rank and resident training year, but not by years in practice. There were site-specific differences based on student sex, but overall no difference.

Conclusions: To our knowledge, this is the first large-scale multi-institutional implementation of a single clinical assessment tool. This study demonstrates the feasibility of a unified approach to clinical assessment across multiple diverse sites. Challenges remain in determining appropriate score distributions and improving consistency in scoring between sites.

Future physicians must be trained to provide safe, high-quality care in every specialty, at every school.[1] Reliable assessment tools are needed to ensure that learners are competent. Despite requirements by regulatory bodies, clinical assessments in undergraduate medical clerkships are imprecise, unreliable, highly variable, lacking in validity evidence, and not comparable between sites, ultimately jeopardizing the quality and safety of patient care.[2–9]

With the shift toward competency-based assessment, multiple initiatives have been aimed at measuring clinical performance.[10,11] Within emergency medicine

From the [1]Department of Emergency Medicine, University of Arizona, Tucson, AZ, USA; the [2]Department of Emergency Medicine, Johns Hopkins University, Baltimore, MD, USA; the [3]Department of Emergency Medicine, East Carolina University, Greenville, NC, USA; the [4]Office of Assessment and Evaluation, Johns Hopkins University, Baltimore, MD, USA; and the [5]Department of Emergency Medicine, University of Washington, Seattle, WA, USA.
Received April 9, 2020; revision received June 8, 2020; accepted June 15, 2020.
The authors have no relevant financial information or potential conflicts to disclose.
Supervising Editor: Daniel P. Runde, MD.
Address for correspondence and reprints: Doug Franzen, MD, MEd; e-mail: franzend@u.washington.edu.
AEM EDUCATION AND TRAINING 2021;5:1–9

doi: 10.1002/aet2.10496

(EM) undergraduate medical education, several large-scale projects have sought to improve the rigor of teaching and assessment, including standardized curricula, examinations, and letters of evaluation.[4,5,7–9] Despite these efforts, there has been no consistent or effective approach to standardizing the tools or processes used in clinical assessment of medical students.

Exacerbating the issue, students applying to EM residencies rotate at multiple sites, where they are assessed using site-specific tools of unknown reliability and validity.[12] Data from these tools are then translated idiosyncratically into other high-stakes products, including grades, Medical Student Performance Evaluations (MSPEs), and Standardized Letters of Evaluation (SLOEs), all vital determinants of whether students are interviewed, ranked, and ultimately matched into EM residency programs.[13–15] Students' career prospects are thus profoundly influenced by site-specific data that are at best imprecise and at worst inaccurate or biased.

While there are significant challenges to standardization of assessment in EM in particular, there are also unique opportunities. Students in EM clerkships are typically assessed on each shift and receive multiple observations from many different assessors, resulting in robust data collection for every student.[12] However, individual assessors must base their assessments on a short observation period, making standardization and reliability particularly crucial. In 2016, in recognition of this opportunity and challenge, a national group of EM education experts collaborated to identify key domains of clinical assessment within EM clerkships and create behaviorally anchored rating scales to measure student proficiency in those domains. The product of this consensus conference was the National Clinical Assessment Tool for Emergency Medicine (NCAT-EM).[16,17] Since then, the NCAT-EM has been adopted by numerous EM clerkships across the country.[18] The purpose of this study is to describe preliminary results of the implementation of the NCAT-EM by a multi-institutional consortium, to inform its continued use in EM clerkships for assessing the clinical competence of medical students.

METHODS

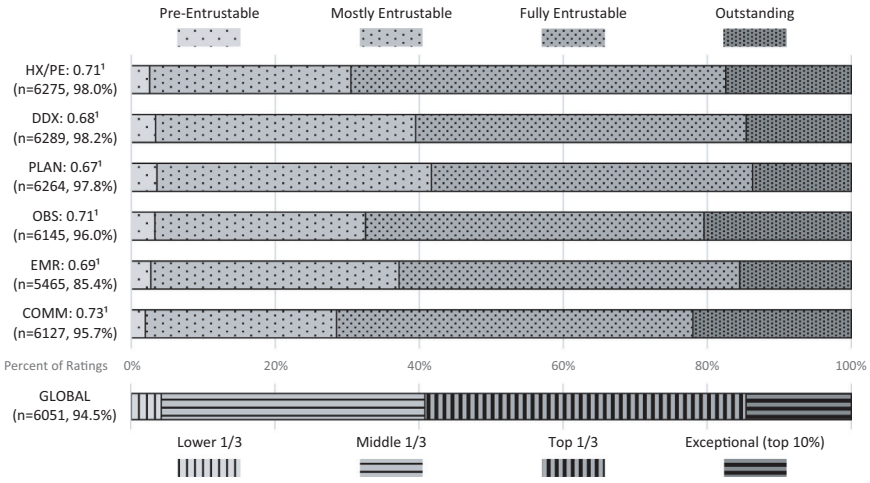
We recruited a convenience sample to participate in this study. Recruitment occurred via e-mail and listservs and through word of mouth at EM education conferences. All U.S. clerkship sites offering an EM rotation for medical students were eligible to participate. Participating sites were required to use the NCAT-EM for clinical assessment of medical students during all ED shifts. Substantive edits to the tool were not permitted. Participating sites were required to enter data for the entire 2017 to 2018 academic year.

We provided assessor instructions to site directors, who were instructed to share them with all assessors at their sites. These instructions included information about how to approach the assessment process and how to correctly use each of the rating scales. They did not provide information about "ideal" score distributions, because all items on the tool except the global rating are criterion referenced. Instructions emphasized the need to assess the student on behaviors directly observed during the clinical shift, with the exception of the global rating question. Site directors were asked to share the instructions with assessors prior to implementation of the NCAT-EM, but the details of how to share these instructions with individual assessors and frequency of review were left to the discretion of site directors. For each student, a unique NCAT-EM was completed at the end of each shift. Each site followed their usual site-specific procedures for assigning, delivering, completing, and tracking assessments, thus avoiding the potentially confounding effect of concurrently changing process variables while a new assessment tool was introduced.

An online, Web-based platform with a backend relational database was created by the Office of Information Technology at the Johns Hopkins University School of Medicine using Microsoft SQL Server. Site directors entered deidentified demographic and assessment data into this database. Individual student codes and assessor codes, autogenerated by the platform, linked individual students and assessors to specific NCAT-EM forms. All data were deidentified prior to aggregation and analysis; site directors alone had access to identifiable data.

We collected student demographic variables including age, sex, training year (M3 vs. M4), home versus visiting rotation, required versus elective/selective rotation, number of prior EM rotations, and whether the student was requesting a SLOE and/or pursuing a career in EM. We collected assessor demographic variables collected including age, sex, academic rank and years in practice (faculty), or postgraduate year (PGY, residents).

The NCAT-EM requires assessment of six domains of clinical performance (Figure 1).[17] Performance in these domains is measured using a four-point rating



[1]Presented as mean rating score where rating (score) was defined as pre-entrustable (0.25), mostly entrustable (0.50), fully entrustable (0.75), and outstanding (1.00).

Figure 1. Clinical rating scale and global score distributions across all assessments.

scale. Scale anchors incorporate EM curricular objectives, ACGME Milestones, and the AAMC EPA descriptors.[8,10,19] The tool also requires assessors to note whether professionalism lapses occurred in any of eight subcategories, with narrative comments required when concerns were identified. A global assessment item requires assessors to stratify learners in a manner similar to the SLOE (i.e., top 10%, top third, middle third, lower third).[4,20] A narrative comment field with the prompt "Please comment on this student's performance today" is included at the end of the tool.[17] The methods by which the tool was derived have been previously described.[16,17]

Data were analyzed by the Office of Assessment and Evaluation at the Johns Hopkins University School of Medicine, using SAS Enterprise Guide 6.1 (SAS Institute, Cary, NC). We report descriptive statistics for demographics and assessment item responses as median and interquartile range (IQR) or frequencies and percentages, as appropriate. We treated rating scale responses as interval-level data with equal distance between each response option. We represented rating scale scores as percentages ("preentrustable" = 25%, "mostly entrustable" = 50%, "fully entrustable" = 75%, and "outstanding" = 100%). Mean clinical rating scale scores represent simple averages of these percentages across all domains and ratings. We used Fisher's exact test for comparison of professionalism and global assessment ratings.

We used analysis of variance (ANOVA) for comparison of mean rating scale scores by site, number of professionalism concerns (0, 1, or 2 or more), and global assessment rating. We used factorial ANOVA to explore differences between student and assessor demographic groups, modeling comparison of mean rating scale scores by site, student or assessor demographic variable, and site-demographic interaction. We further explored significant findings with follow-up pairwise comparisons using Sidak adjusted p-values. We estimated internal consistency using Cronbach's coefficient alpha. This project was reviewed by the University of Arizona Institutional Review Board and all consortium member sites, either independently or through a formal institutional agreement with the University of Arizona.

RESULTS

Thirteen sites were ultimately able to meet requirements to participate, representing 12 states and all geographic regions. These sites reported data for all students rotating during the entire 2017 to 2018 academic year, entering 6,402 NCAT-EM assessments for 748 students by 704 assessors. On average, each student received 8.6 assessments, but there was considerable variability in this figure, ranging from 4.3 to 11.3 (Table 1). More than three-quarters of students received seven or more assessments, but only 7% received more than 12 (Table 2). On the assessor side, more than half completed five or fewer assessments, whereas 13% completed 20 or more (Table 3).

Demographic data for both students and assessors are presented in Figure 2. Demographic variables were unavailable for a substantial proportion of assessments.

Table 1
Number of Students, Assessors, and Assessments by Site

Site	Total No. of Students	Total No. of Assessors	Mean No. of Assessments/ Student	Mean No. of Assessments/ Assessor	Total No. of Assessments	Mean Clinical Rating Scale Score
East Carolina University	102	53	10.8	20.7	1,098	65.3%
Harvard	29	48	7.0	4.2	203	70.2%
Kaweah Delta	45	16	4.3	12.2	195	66.4%
Michigan State University	41	71	7.3	4.2	298	65.7%
Regions Medical Center	31	25	8.9	9.0	278	59.6%
Reading Hospital	39	49	11.3	9.0	442	78.2%
Texas Tech, El Paso	86	39	8.5	18.7	731	74.0%
University of Arizona	80	87	10.8	10.0	867	69.6%
University of Colorado	64	96	7.8	5.2	501	78.8%
University of Florida, Jacksonville	94	55	8.2	14.0	773	72.1%
University of Missouri	30	39	4.6	3.5	137	69.9%
University of Nevada, Las Vegas	55	58	8.9	8.4	487	70.0%
University of Pennsylvania	52	68	7.5	5.8	392	58.5%
Total	748	704	8.6	9.1	6,402	

Table 2
Number of Assessments per Student

Number of Assessments	Number of Students	Percentage of Students
1–3	44	5.88
4–6	156	20.86
7–9	255	34.09
10–12	236	31.55
13–15	44	5.88
16–18	4	0.53
19–21	4	0.53
22–24	5	0.67
Total	748	100.00

Table 3
Number of Assessments per Assessor

Number of Assessments per Assessor	Number of Assessors	Percentage of Assessors
1–5	382	54.26
6–10	142	20.17
11–15	49	6.96
16–20	38	5.40
21–25	32	4.55
26–30	21	2.98
31–35	23	3.27
≥36	17	2.41
Total	704	100.00

Among the students for whom data were available, the vast majority were in the fourth year of medical school. Slightly more than half were male. Two-thirds were assessed in elective or selective rotations. Slightly more than half were planning residency training in EM. Among the assessors for whom data were available, slightly more than half were male, and nearly three-quarters were faculty, most of whom were at the instructor or assistant professor level. The largest proportion of residents completing assessments were in the PGY3 year, although PGY2 and PGY4 residents also commonly served as assessors.

Rating scale and global assessment scores are summarized in Figure 1. Data were recorded for 95.2% of all rating scale prompts. The anchors "mostly entrustable" (n = 11,723, 30.5%) and "fully entrustable/Milestone 1" (n = 17,456, 45.4%) were the most frequently used. "Outstanding/Milestone 2" was used less often (n = 6,323, 16.5%). "Preentrustable" was rarely used, accounting for less than 3% of ratings (n = 1,063, 2.8%). Most missing data were due to "unable to assess" (n = 1,527, 82.7%) rather than lack of response (n = 320, 17.3%). Over half of the "unable to assess" responses were in "emergency recognition and management" (n = 872, 57.8%).

Mean scores were similar across the six clinical rating scale domains, ranging from 67.1% to 72.9%. The domains with slightly higher mean scores were "Focused history and physical exam skills (71.1%)"; "Observation, monitoring and follow-up (71.2%)"; and "Patient and team-centered communication (72.9%)." The domains with slightly lower scores were "Ability to generate a prioritized differential diagnosis

Variable	Participants with available data N (%)	Result based on available data				
Student Data						
Age	162 (21.7%)	Median (Years)	IQR			
		28	27-30			
Sex	586 (78.3%)	Male	Female			
		335 (57.1%)	251 (42.9%)			
Training level	580 (77.5%)	M3	M4			
		34 (5.9%)	546 (94.1%)			
Rotation type	580 (77.5%)	Required	Elective/selective			
		207 (35.7%)	373 (64.3%)			
Number prior EM rotations	294 (39.3%)	0	1	2 or more		
		132 (44.9%)	119 (40.5%)	43 (14.6%)		
Residency training plan	571 (76.3%)	EM	Non-EM			
		315 (55.2%)	256 (44.8%)			
Assessor Data						
Age	441 (56.3%)	<30	30-39	40-49	50-59	>59
		34 (4.8%)	189 (26.9%)	116 (16.5%)	39 (5.5%)	19 (2.7%)
Sex	645 (85.4%)	Male	Female			
		382 (63.6%)	219 (36.4%)			
Faculty vs. resident	595(84.5%)	Faculty/Fellow	Resident			
		422 (71.3%)	170 (28.6%)			
Faculty rank	422 (100% of those designated as faculty	Fellow	Instructor	Assistant	Associate	Professor
		3 (0.7%)	99 (23.5%)	225 (53.3%)	67 (15.9%)	28(6.6%)
Resident training year	170 (100% of those designated as residents)	PGY1	PGY2	PGY3	PGY4-5	
		10 (5.9%)	50 (29.4%)	73 (42.9%)	37 (21.8%)	
Assessor practice years	298 (42.5%)	1-5	6-15	16-25	>25	
		113 (37.9%)	110 (36.9%)	50 (16.8%)	25 (8.4%)	

Figure 2. Student and assessor demographic data. All results are expressed as number of NCAT assessments within each category, followed by percent of assessments *for which data are available*. Three assessors had both resident training year and faculty rank listed in the data set; these three assessors were eliminated from those analyses. NCAT = National Clinical Assessment Tool.

(67.9%)"; "Ability to formulate a plan (67.1%)"; and "Emergency recognition and management (68.9%)." Rating scale data for each domain are presented in Figure 1.

Professionalism concerns were noted in <1% of all assessments, with 88 total concerns noted on 62 (0.15%) discrete assessments, accounting for 46 students (6.1%). Every professionalism subcategory was used at least once. Students with assessments reflecting one or more professionalism concerns had significantly lower mean rating scale scores than students without any professionalism concerns ($p < 0.0001$). The mean rating scale score for students with no professionalism concerns was 69.7%, compared to 55.2% for students with one concern and 42.8% for students with two or more concerns. Professionalism data are analyzed in greater detail elsewhere.[21]

Global assessment data were available for 94.5% of all assessments. There was a positive skew, with only 256 (4.0%) of all students rated at the "lower third" level. Lower global assessment scores were strongly associated with lower mean rating scale scores ($p < 0.0001$) and professionalism concerns (as above). Students rated "lower third" had a mean rating scale score of 43.6%, compared to 59.5% for those rated "middle third," 73.9% for those rated "upper third," and 89.9% for those rated "top 10%."

Overall internal consistency in NCAT-EM scores was high within each participating institution. Every site had a Cronbach's coefficient alpha >0.80 (range = 0.81 to 0.91). There were, however, significant differences *between* sites in mean clinical rating scale scores, with mean scores ranging from 58.5% to 78.8% ($p < 0.0001$, Table 1).

Factorial ANOVA tests exploring differences in assessment scores based on demographic variables are presented below. All factorial ANOVAs were conducted with limited demographic data, which many sites had difficulty consistently collecting and reporting. All analyses are based on the subset of assessments for which demographic data were available (Figure 2).

Sex

The majority of both students and assessors were men (Table 3). There were no statistically significant differences in rating scale scores by student sex overall ($p = 0.17$). However, there were differences by site ($p < 0.0001$) and there was a significant interaction between student sex and site ($p < 0.0001$), meaning that there were systematic score differences by student sex at some sites. There were four sites with differences by student sex, with women scoring slightly higher than men at three of these.

Regarding the effect of assessor sex, overall there were no statistically significant differences in scores assigned by assessors based on their sex ($p = 0.31$). However, differences by site ($p < 0.0001$) and a significant interaction between assessor sex and site ($p < 0.0001$) were seen. This was noted at six sites, with female assessors assigning slightly higher scores than males at three of these sites and the reverse at the other three. Two sites were excluded from both sex analyses due to a lack of data.

Student Training

The vast majority of students were assessed on a fourth-year rotation (Figure 2). There were significant differences in rating scale scores by student training year, with M3 students averaging 63.0% and M4 students averaging 67.0% ($p = 0.002$). Three sites were not included in this analysis due to a lack of data.

There were no statistically significant differences in rating scale scores by number of prior EM rotations ($p = 0.37$), but there were differences by site ($p < 0.0001$) and an interaction between number of prior EM rotations and site ($p = 0.03$). At one site, students who had completed one or more prior EM rotation had significantly higher mean scores. This effect was not consistent or significant across all sites. Five sites were not included in this analysis due to a lack of data.

There were also significant differences in rating scale scores for EM-bound students compared to those pursuing careers in other specialties ($p = 0.003$). The overall mean rating scale score for EM-bound students was 72.4%, versus 69.2% for non–EM-bound students. This relationship was significant at the institutional level for only one site. Four sites were not included in this analysis due to a lack of data.

Assessor Training and Experience

Overall, faculty gave slightly lower scores than residents (68.53 vs. 70.50, $p < 0.0001$). There were significant differences in rating scale scores assigned by faculty based on academic rank ($p = 0.007$). Mean score assigned by clinical instructors was 64.7%, compared to 69.9% for assistant professors, 69.2% for associate professors, and 66.9% for professors. Follow-up ANOVA with pairwise comparisons indicated that both assistant and associate professors scored students significantly higher than clinical instructors ($p < 0.0001$ for both pairwise comparisons). Four sites were not included in this analysis due to a lack of data. There was no statistically significant difference in scores by faculty years in practice ($p = 0.22$).

For resident assessors, there were small but statistically significant differences in rating scale scores by training year ($p < 0.0001$). Mean score assigned by PGY1 residents was 72.5%, compared to 70.1% for PGY2, 67.8% for PGY3, and 77.4% for PGY4 and -5. Five sites were not included in this analysis due to a lack of data.

DISCUSSION

This study represents the first large-scale multicenter aggregation of clinical assessment data for medical students in EM clerkships. It is proof of concept that standardization of assessment tools and comparison of data across institutions is feasible. It also illustrates several important points to be addressed in future studies.

Regarding the tool itself, all categories of performance were used. Data from the clinical domains yielded a positively skewed score distribution similar to that previously described for other EM assessment tools.[22,23] The lowest category ("preentrustable") was rarely used, and the top category ("outstanding/Milestone 2") was used much more often (2.8% vs. 16.5%). "Fully entrustable" was also used more often than "mostly entrustable" (45.4% vs. 30.5%). This observation could be explained by the preponderance of M4 students in our sample. As senior learners, they

would be expected to have greater mastery of the knowledge and skills reflected in the domains on the NCAT-EM compared to more junior learners. However, it is also possible that this finding simply reflects current assessment culture, with a tendency of raters to focus their scores at the higher end of the range.

The latter hypothesis is supported by the significant variability in mean clinical rating scores between institutions. It is unlikely that the large variation we observed in mean rating scale scores across sites (ranging from 58% to 79%) is explained entirely by differences in student performance. It is more likely that this variability is related to differences in assessment culture across these settings. "Culture" is a complex construct, but in this case may include the grading and assessment philosophies of thought leaders at each institution, explicit or implicit expectations from institutional leadership regarding grade distribution, specific experiences within each clerkship regarding grading and assessment, and individual assessor tendencies.

This observation speaks strongly to the need for rater training and calibration. Assuming that classes of medical students are relatively comparable in their clinical performance across institutions, we would expect that an "average" student at one school would earn roughly the same "average" score at another school. However, our data demonstrate that this "average" student may score 79% at one school but only 58% at another. If we hope to be able to compare students meaningfully across institutions, we need to develop a shared mental model for levels of proficiency and a common system for applying them in clinical assessment. This could potentially be accomplished through a more robust approach to assessor development. The impact of such a program on interinstitutional reliability merits further study. The ability to meaningfully compare students across institutions becomes increasingly important with the recent announcement of Step 1 changing to pass/fail.[24]

Professionalism lapses were rare, but every subdomain was used at least once, indicating a need to retain all subcategories. The strong correlation between professionalism lapses and low scores merits further investigation. This finding may be explained in part by the "halo/horns effect," wherein an assessor's strong reaction to one aspect of a student's performance colors their perception of other aspects of that student's performance. This is another area where rater training and calibration may help raters better distinguish between different dimensions of student performance and score them accordingly. It should be noted that the correlation between professionalism lapses and low scores in other domains persisted across all assessors—not just the one or two who noted the lapse. This finding may suggest that *any* observed unprofessional behavior is a harbinger for other performance problems and should be taken seriously as a target for further investigation and remediation.

The strong correlation between rating scale scores and global assessment score is expected, as students who demonstrate strong clinical skills are more likely to be perceived as excellent in other domains not explicitly assessed. However, clinical domains on the NCAT-EM are criterion referenced, meaning that if 100% of the students demonstrate the behaviors described in the "outstanding" anchor for a given domain, then all of those students can and should be scored as "outstanding" in that domain. There is thus no expected score distribution for the clinical domains. The global assessment item, by contrast, is norm referenced, specifically asking raters to rate the student's performance relative to a group of their peers and describing an expected score distribution. However, in our data set, assessors only placed 4% of students in the lower third, while the top third and top 10% ratings were assigned to more than half of students, a state that is obviously mathematically infeasible but similar to the global rating results reported in one study of the SLOE.[25]

The analysis by student training year was limited by the small number of M3 students in our sample. Our data did demonstrate higher scores for M4 students compared to M3s, which is expected as students should improve as they progress along the learning continuum. However, the magnitude of this difference was very small, and it is unclear whether it is meaningful. Additional studies are needed to delineate these differences and changes over time.

There was also a small difference between the scores for EM-bound students compared to non–EM-bound students, with higher scores for the EM-bound group. This is an expected finding, as students typically invest greater effort and thus perform better during rotations within their chosen specialty. Unexpectedly, no difference in scores was observed based on the number of prior EM rotations completed. This finding may be related to the adverse impact of rotating at an unfamiliar institution on

clinical performance. It may also be due to differences in assessment culture between institutions. These analyses were hampered by missing demographic data, and further study is needed to explore how these variables affect performance.

There was a large amount of interinstitution variability, suggesting differences in curriculum, assessor development activities, and expectations for students, both explicit and implicit. More work is necessary to improve consistency in both articulation of expected student performance and standardization of assessor development. Further work in comparing NCAT-EM scores generated in institutions with similar curricula (i.e., M3, M4, elective, mandatory) could be done with a larger consortium of sites.

While there was no evidence of sex bias overall, there were unexpected differences in scores by sex at some institutions—a finding that is obviously concerning. If these findings persist over time, investigation into sources of bias will be needed, as will steps to remedy disparities.

We found small differences in ratings based on assessor rank and experience, with faculty assigning slightly lower scores than residents. We did not find a clear systematic association between assessment scores and resident program year, academic rank, or years in practice. We found much larger differences in assessment scores between institutions than we did between assessor groups, suggesting that institutional assessment culture is a stronger predictor of rater behavior than experience.

LIMITATIONS

Although no data were systematically excluded from analysis, there is no way to capture whether any assessments might have been lost or inadvertently omitted from the data set. While our sample was large, it included only a small subgroup of M3 students, making descriptions of score distribution and other metrics for more junior learners unfeasible. This may also have biased our comparisons of M3 versus M4 students. Our sample also included a large proportion of EM-bound students and elective rotations. Therefore, our results may not be generalizable to students in required rotations. We did not collect data on completion rates or timeliness of assessment completion. The reporting of demographic data varied substantially between sites, with large amounts of missing demographic data that may have biased our results. Finally, the recruitment process for this study required a high level of engagement for site directors, coordinators, and assessors. Results may not generalize to settings where clerkship directors, faculty, and staff are less actively engaged in the learner assessment process.

CONCLUSIONS

In summary, we found that multi-institutional implementation of a standardized, consensus-derived, specialty-specific clinical assessment tool for medical students is feasible. This approach holds great promise for improving the reliability, validity, and comparability of emergency medicine assessment across institutions, although further work in this regard is needed. Future studies should focus on measuring the impact of assessor training and calibration on interinstitutional reliability. Optimal score distributions must be defined, and further study will be needed to assess the impact of training year, specialty choice, and prior rotations on scores. Future studies should seek to evaluate the validity of National Clinical Assessment Tool for Emergency Medicine data and to correlate National Clinical Assessment Tool for Emergency Medicine scores with other external measures of performance. Ultimately, we believe that use of a standardized clinical assessment tool and centralized data collection process will provide valuable information to medical schools and residency programs and help to assure a competent physician workforce.

References

1. Marceau M, Gallagher F, Young M, St-Onge C. Validity as a social imperative for assessment in health professions education: a concept analysis. Med Educ 2018;52:641–53.
2. Liason Committee on Medical Education. Functions and structure of a medical school: standards for accreditation of medical education programs leading to the MD degree. Available at: http://lcme.org/wp-content/uploads/filebase/standards/2019-20_Functions-and-Structure_2018-09-26.docx. Accessed Jan 18, 2020.
3. Alexander EK, Osman NY, Walling JL, Mitchell VG. Variation and imprecision of clerkship grading in U.S. medical schools. Acad Med J Assoc Am Med Coll 2012;87:1070–6.
4. Keim SM, Rein JA, Chisholm C, et al. A standardized letter of recommendation for residency application. Acad Emerg Med 1999;6:1141–6.
5. Senecal EL, Heitz C, Beeson MS. Creation and implementation of a national emergency medicine fourth-year student examination. J Emerg Med 2013;45:924–34.

6. Love JN, Smith J, Weizberg M, et al. Council of Emergency Medicine Residency Directors' standardized letter of recommendation: the program director's perspective. Acad Emerg Med 2014;21:680–7.
7. National Board of Medical Examiners. Emergency Medicine Advanced Clinical Exam. Available at: https://www.nbme.org/Schools/Subject-Exams/Subjects/ace_emergmed.html. Accessed Jan 18, 2020.
8. Manthey DE, Ander DS, Gordon DC, et al. Emergency medicine clerkship curriculum: an update and revision. Acad Emerg Med 2010;17:638–43.
9. Tews MC, Wyte CM, Coltman M, et al. Developing a third-year emergency medicine medical student curriculum: a syllabus of content. Acad Emerg Med 2011;18(Suppl 2):S36–40.
10. American Association of Medical Colleges. Core Entrustable Professional Activities for Entering Residency: Faculty and Learners' Guide. Available at: https://store.aamc.org/downloadable/download/sample/sample_id/66/%20. Accessed Apr 3, 2020.
11. Chen HC, McNamara M, Teherani A, Cate OT, O'Sullivan P. Developing entrustable professional activities for entry into clerkship. Acad Med 2016; 91:247–55.
12. Lawson L, Jung J, Franzen D, Hiller K. Clinical assessment of medical students in emergency medicine clerkships: a Survey of Current Practice. J Emerg Med 2016;51:705–11.
13. Katzung KG, Ankel F, Clark M, et al. What do program directors look for in an applicant? J Emerg Med 2019;56:e95–101.
14. Crane JT, Ferraro CM. Selection criteria for emergency medicine residency applicants. Acad Emerg Med 2000;7:54–60.
15. Breyer MJ, Sadosty A, Biros M. Factors affecting candidate placement on an emergency medicine residency program's rank order list. West J Emerg Med 2012;13:458–62.
16. Hiller KM, Franzen D, Lawson L, et al. Clinical assessment of medical students in the emergency department, a national consensus conference. West J Emerg Med 2017;18:82–3.
17. Jung J, Franzen D, Lawson L, et al. The National Clinical Assessment Tool for medical students in the emergency department (NCAT-EM). West J Emerg Med 2018;19:66–74.
18. Hiller KM. NCAT-EM: Update on year 2 of use. Presented at the Council of Residency Directors in Emergency Medicine; Seattle, WA; April 1, 2019.
19. Beeson MS, Holmboe ES, Korte RC, et al. Initial validity analysis of the Emergency Medicine Milestones. Acad Emerg Med 2015;22:838–44.
20. Love JN, Ronan-Bentle SE, Lane DR, Hegarty CB. The standardized letter of evaluation for postgraduate training: a concept whose time has come? Acad Med 2016;91:1480–2.
21. Emery M, Parsa MD, Watsjold BK, Franzen D. Assessment of professionalism during the emergency medicine clerkship using the national clinical assessment tool for medical students in emergency medicine. Acad Emerg Med 2020;27.
22. Love JN, Deiorio NM, Ronan-Bentle S, et al. Characterization of the Council of Emergency Medicine Residency Directors' standardized letter of recommendation in 2011–2012. Acad Emerg Med 2013;20:926–32.
23. Grall KH, Hiller KM, Stoneking LR. Analysis of the evaluative components on the Standard Letter of Recommendation (SLOR) in emergency medicine. West J Emerg Med 2014;15:419–23.
24. United States Medical Licensing Examination. InCUS – Invitational Conference on USMLE Scoring. Change to pass/fail score reporting for Step 1. Available at: https://www.usmle.org/InCUS/. Accessed Apr 3, 2020.
25. Jackson J, Bond M, Love J, Hegarty C. Emergency medicine standardized letter of evaluation (SLOE): findings from the new electronic SLOE format. J Grad Med Educ 2019;11:182–6.