Published in final edited form as:
Acad Med. 2022 April 01; 97(4): 544–551. doi:10.1097/ACM.0000000000004222.

Workplace-Based Entrustment Scales for the Core EPAs: A Multisite Comparison of Validity Evidence for Two Proposed Instruments Using Structured Vignettes and Trained Raters

Michael S. Ryan, MD, MEHP [professor and assistant dean],
Clinical Medical Education, Department of Pediatrics, Virginia Commonwealth University, Richmond, Virginia

Asra R. Khan, MD [associate professor, director, Doctoring and Clinical Skills course, and clerkship director],
Department of Internal Medicine, University of Illinois College of Medicine, Chicago, Illinois

Yoon Soo Park, PhD [director],
Health Professions Education Research, and member of the faculty, Harvard Medical School and Massachusetts General Hospital, Boston, Massachusetts

Cody Chastain, MD [assistant professor],
Department of Internal Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee.

Carrie Phillipi, MD, PhD [professor and vice chair of education],
Department of Pediatrics, Oregon Health & Science University, Portland, Oregon.

Sally A. Santen, MD, PhD [professor and senior associate dean],
Assessment, Evaluation, and Scholarship, Department of Emergency Medicine, Virginia Commonwealth University, Richmond, Virginia.

Beth A. Barron, MD [associate professor and associate director],
Simulation, Department of Internal Medicine, Columbia University School of Medicine, New York, New York.

Vivian Obeso, MD [associate professor and assistant dean],
Curriculum and Medical Education, Department of Internal Medicine, Florida International University, Miami, Florida.

Sandra L. Yingling, PhD [assistant professor and associate dean],
Educational Planning and Quality Improvement, Department of Medical Education, University of Illinois College of Medicine, Chicago, Illinois;

Core Entrustable Professional Activities for Entering Residency Pilot Program

Correspondence should be addressed to Michael S. Ryan, 1201 E. Marshall St., Suite 4-200, Box 980565, Richmond, VA 23298-0565; telephone: (804) 828-4589; michael.ryan1@vcuhealth.org; Twitter: @MichaelSRyanMD.

Other disclosures: None reported.

Ethical approval: This study was deemed exempt by the University of Illinois College of Medicine Institutional Review Board on May 17, 2019 (IRB No. 2019–0504).

Previous presentations: Content from this report was presented at the 2020 Association of American Medical Colleges Learn Serve Lead Annual Meeting (November 2020; virtual).

Abstract

Purpose—In undergraduate medical education (UME), competency-based medical education has been operationalized through the 13 Core Entrustable Professional Activities for Entering Residency (Core EPAs). Direct observation in the workplace using rigorous, valid, reliable measures is required to inform summative decisions about graduates' readiness for residency. The purpose of this study is to investigate the validity evidence of 2 proposed workplace-based entrustment scales.

Method—The authors of this multisite, randomized, experimental study used structured vignettes and experienced raters to examine validity evidence of the Ottawa scale and the UME supervisory tool (Chen scale) in 2019.

The authors used a series of 8 cases (6 developed de novo) depicting learners at preentrustable (less-developed) and entrustable (more-developed) skill levels across 5 Core EPAs. Participants from Core EPA pilot institutions rated learner performance using either the Ottawa or Chen scale. The authors used descriptive statistics and analysis of variance to examine data trends and compare ratings, conducted interrater reliability and generalizability studies to evaluate consistency among participants, and performed a content analysis of narrative comments.

Results—Fifty clinician-educators from 10 institutions participated, yielding 579 discrete EPA assessments. Both Ottawa and Chen scales differentiated between less- and more-developed skill levels ($P < .001$). The interclass correlation was good to excellent for all EPAs using Ottawa (range, 0.68–0.91) and fair to excellent using Chen (range, 0.54–0.83). Generalizability analysis revealed substantial variance in ratings attributable to the learner–EPA interaction (59.6% for Ottawa; 48.9% for Chen) suggesting variability for ratings was appropriately associated with performance on individual EPAs.

Conclusions—In a structured setting, both the Ottawa and Chen scales distinguished between preentrustable and entrustable learners; however, the Ottawa scale demonstrated more desirable characteristics. These findings represent a critical step forward in developing valid, reliable instruments to measure learner progression toward entrustment for the Core EPAs.

Competency-based medical education (CBME) requires well-defined, rigorous, valid, evidence-based measures of competence. [1–5] Faculty members and residency program directors agree that a mutually understood standard of competence would improve the transition of medical students to their role as interns. [6–11] Until recently, medical schools have lacked uniformity in their measurements of students' competence in key clinical skills at graduation. To meet that need, in 2014, the Association of American Medical Colleges (AAMC) published the Core Entrustable Professional Activities for Entering Residency (Core EPAs). The Core EPAs are a set of 13 integrated clinical activities that medical students may be entrusted to perform under indirect supervision upon entering residency. [12] The AAMC selected 10 undergraduate medical education (UME) institutions at which to implement medical student CBME by applying this Core EPA framework. [13]

Performance of the Core EPAs in authentic workplace settings is central to the concepts of competence and entrustment. [6,14] In an optimal workplace-based assessment (WBA) system, learners would receive multiple formative assessments of their performances, and the feedback from those assessments would help them identify areas for further development.[15,16] To make entrustment decisions at the end of training, medical schools need reliable, valid WBAs. Currently, no single framework serves as a gold standard for formative assessment of the Core EPAs, and validity evidence for EPA entrustment scales is limited.[17]

While the literature includes descriptions of numerous entrustment scales, 2 have been proposed for use in the UME setting. One comes from a scale found on 2 similar instruments: the Ottawa Clinic Assessment Tool (OCAT)[18] and the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE).[19] This scale (hereafter, the Ottawa scale) is retrospective in that it asks the clinical supervisor how much intervention the learner needed for the observed activity.[19] Another scale developed by Chen and colleagues (hereafter, the Chen scale)[20] asks the clinical supervisor to identify the amount of supervision the learner will need in the future. See Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/B139.

Early in the work of the Core EPA pilot, a task force considered the merits of these 2 scales as well as others. The task force and pilot participants did not achieve consensus, and, as a result, the Core EPA pilot schools incorporated the Ottawa scale, the Chen scale, or both into their EPA-based WBAs.[21]

While individual programs have incorporated both the Ottawa and Chen scales for formative assessment of medical students, the comparative validity of these scales in the same setting has yet to be reported.[22] One common method used to assess the validity of performance-based assessment tools is Messick's framework, which examines 5 sources of validity: content, response process, internal structure, relationship to other variables, and consequences.[23] In this study, we assessed the performance of the 2 scales through the lens of Messick's framework using vignettes.

Calaman and colleagues recently described the development of scripted standard-setting videos to represent levels of competence for 1 Core EPA: patient handovers (Core EPA 8).[24] Through these vignettes, clinical faculty may become better prepared to rate the unscripted, real-world performances of the learners they observe.[24] Similarly, we created a series of 6 structured vignettes, each including a learner performing 1 or more of the Core EPAs (see Table 1). We asked faculty from the Core EPA pilot institutions to rate the performance of the standardized learner in each of these vignettes and in 2 other, existing vignettes,[25] using either the Ottawa or Chen scale. We aimed to apply Messick's framework[23] to examine the validity (specifically, the content, response process, and internal structure validity evidence) of the Ottawa and Chen scales for use in WBAs in UME. Based on the language used within the anchors of the Ottawa scale, we hypothesized that the Ottawa scale would be deemed more intuitive to clinical supervisors and, thus, possess more desirable characteristics compared with the Chen scale.

Method

We conducted a multisite, randomized, experimental study using a total of 8 structured vignettes to examine the Ottawa and Chen scales. We created 6 new vignettes to represent a single, discrete, direct observation in a simulated patient care activity. Our goal was to use these vignettes and 2 existing vignettes [25] to explore the relative strengths of each scale in discriminating between medical student performance at the preentrustable (less-developed) level and the entrustable (more-developed) level (see Table 1). [12] We reasoned that if we observed discrimination between these 2 general levels of performance, then further studies could investigate each scale's ability to delineate more discrete levels of performance.

Creation of vignettes

Using a process similar to that described by Calaman and colleagues, [24] we developed, de novo, 6 vignettes, each depicting a single encounter with a learner. First, we used behavioral anchors for learner development provided in the Core EPA toolkits [21] to serve as a blueprint for case vignettes. We then created 2 scripts for each EPA. One depicted a student early in skill development (less-developed), and the second depicted a student at more advanced stages of skill development (more-developed). Because we created scripts based on the anchors provided in the Core EPA toolkits, we did not design the cases to achieve a particular score with respect to either scale.

Three authors (A.R.K., M.S.R., and S.L.Y.) reviewed drafts of scripts. Rather than create vignettes for Core EPA 8 (Handover), we included 2 previously developed video vignettes [25] that depicted that activity. For the remaining 4 EPAs, we developed 4 new videos and 2 written notes (see Table 1). We recruited a fourth-year medical student to portray the student in all vignettes, 1 author (A.R.K.) to portray the preceptor in all vignettes, and a standardized patient to portray the patient in all vignettes. One author (A.R.K.) drafted the note used in each of the 2 written note vignettes. All authors helped revise these notes. See Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/B139.

The authors recognized that, in a realworld setting, each individual vignette (e.g., oral presentation, written note, physical examination) could serve as an opportunity to rate more than 1 EPA. Given this, we asked raters to assess multiple EPAs for some vignettes.

Setting and participants

Faculty members from the Core EPA pilot institutions who were either members of the pilot team or actively engaged in implementation at their institution served as participants. We targeted this population because pilot institutions were required to implement Core EPAs within their curriculum, and, thus, we anticipated they would be most familiar with the practical application of the EPAs.

We obtained informed consent from each participant using a standard template. We conducted the study at a prescheduled meeting of Core EPA pilot team members. We integrated the study into the preexisting agenda; therefore, all attendees received the content. We also conducted the study at Core EPA meetings held at each institution. In analysis, we excluded any participants who indicated they had no clinical responsibilities because we felt their impressions may not represent that of front-line clinical faculty. To maximize consistency in the rating process, one of us (M.S.R.) created a private YouTube channel that included detailed instructions, video vignettes, and links to rating forms. All items available on the channel were shown in the same sequence at each rating session.

Randomization

Participants were randomized to either the Ottawa or the Chen scale using a stratified sampling technique as follows: We first identified the scale(s) used at each institution. Based on those findings, participants were stratified into groups according to the scales they were already familiar with (i.e., Ottawa scale only, Chen scale only, both scales, neither scale). Next, a third party at the AAMC randomized participants to either the Ottawa or Chen group, ensuring balanced distribution. We randomized the order of vignettes a priori but then, as mentioned, maintained this sequence throughout the study.

Data collection

We used a Qualtrics survey (Provo, Utah) to gather the following information from participants:

1. Demographic characteristics (institution, specialty, gender, prior experience with EPA-based WBAs, and experience with learners in years);
2. Their postvignette ratings and narrative explanations for each rating; and
3. Their reflections on the advantages and disadvantages of the scale to which they were assigned.

Before showing the participants the vignettes, we collected their demographic information. Participants completed their ratings after they read or viewed each vignette. Participants then provided a narrative explanation of their rating choice. Finally, participants were asked to comment on the advantages or disadvantages of the scale to which they were randomized. The study comprised a single, long survey, so participants could go back and change answers until they submitted the tool. The survey took about 45 minutes to complete.

Analysis

Ratings on each scale were converted to a numerical score (1 to 5, 1 representing the lowest score and 5 representing the highest) to aid interpretation. We used descriptive statistics and analysis of variance to examine trends in the data, and we used *t* tests and analysis of variance to compare differences in ratings between scales and between learner entrustment levels across vignettes. We examined interrater reliability (IRR) based on intraclass correlations (ICCs) and conducted generalizability studies.[26] To evaluate consistency among participants, to examine factors contributing to variability in scores, and to estimate the standardized error of measurement between scales, we used a fully crossed design as follows: learner (*p*) × rater (*r*) × EPA (*e*). We considered performance in a given EPA similarly to how raters typically consider performance on individual stations in an observed structured clinical examination (OSCE): namely, that performance in 1 station may be correlated with performance in another (e.g., history taking may be associated with oral presentation skills), but that the observations are independent of each other. Therefore, we not only examined the overall generalizability of the scales across EPAs but also considered the EPA itself to function as a facet. Two of us (M.S.R. and S.L.Y.) independently conducted content analyses[27] to identify themes in narrative comments. We conducted all data compilation and analyses using Stata 16 (College Station, Texas). The institutional review board of the University of Illinois at Chicago approved this study.

Results

A total of 63 participants rated Core EPA-related skills. After excluding blank or limited responses (n = 8) and participants who had never practiced clinically (n = 5), we included data from 50 participants in the analysis. Of these, 22 were assigned to the Ottawa scale and 28 were assigned to the Chen scale. Table 2 provides a summary of demographic data from the study participants.

Participants completed a total of 579 EPA-based assessments. Raters in both groups were able, for each EPA, to differentiate between less- and more-developed EPA-related skill levels. For less-developed skill levels, raters using the Ottawa scale demonstrated less variability in scores across EPAs compared with raters using the Chen scale (mean standard deviation for scores on the Ottawa scale across EPAs ranged from 0.46 to 0.93; mean standard deviation for scores on the Chen scale across EPAs ranged from 0.81 to 0.99). A summary of descriptive statistics by scale and learner level is provided in Table 3.

Interrater reliability

The 2 scales were assessed for IRR using intraclass correlation (ICC) and 95% confidence intervals. [28] The Ottawa scale demonstrated good to excellent IRR for all 5 EPAs included in the study (ICC = 0.68–0.91). The Chen scale demonstrated good to excellent IRR for EPA 1 (History and physical), EPA 6 (Oral presentation), and EPA 8 (Handover) (ICC = 0.67–0.83) but only fair IRR for the 2 EPAs assessed using a written note (EPA 2 [Differential diagnosis] and EPA 5 [Encounter note]); the ICCs for these were, respectively, 0.54 and 0.58. A summary of IRR is provided in Table 4.

Generalizability study

We then conducted a generalizability study to describe the variance, reliability, and standard error of measurement of the 2 scales. The variance attributed to the learner was similar for both scales (18.5% for Ottawa vs 18.2% for Chen); however, Ottawa had, compared with Chen, less residual variance (17.8% vs 24.5%) and less rater variability (2.3% vs 6.1%). The Ottawa scale had better sensitivity to differentiate scores among the Core EPAs because variance attributed to the individual EPA (learner × EPA) was greater for the Ottawa scale (59.6%) compared with the Chen scale (48.9%). Reliability was similar for both scales (0.645 for Ottawa vs 0.684 for Chen), while the relative standard error of measurement was greater for the Ottawa scale (0.479) than for the Chen scale (0.418). A summary of variance components and reliability derived from the generalizability study is provided in Chart 1.

Qualitative analysis

Content analysis revealed 4 key themes: challenges applying the scales to direct observation, the nature of entrustment decisions, limitations of using the scales in a simulated environment, and the distinct advantages of each scale.

Challenges applying scales to direct observation.

A key issue was the “fit” of each scale, particularly for ratings requested in the absence of direct observation (e.g., Core EPA 5 [Encounter note]). As one participant stated, “[I] would never prevent [a] student from performing this activity [Core EPA 5]. So, you have to give it 'stakes' [for the supervisor] such as it will go in the chart without review and editing." In essence, participants mentally revised the anchors to make the scale fit with their observations.

Nature of entrustment decisions.

A second theme involved the interactive, interpersonal nature of supervision and entrustment decisions. One participant stated, "prospective entrustment is more colored by *how trusting a preceptor is*, even though the student may have demonstrated competence." The participant explained that even if a student performed well, the supervisor may not trust the learner due to experiences unrelated to the student.

Limitations of using scales in a simulated environment.

Clinical supervision is a highly interactive process in which supervisors create and observe "educational differentials" through dialog with trainees. Participants noted the limitations of using entrustment scales in a simulated setting, including the fact that the supervisor in the video vignettes made different decisions than they may have made in the situation ("I would have steered the student earlier"). Participants also noted that applying the scales to hypothetical situations "added a layer of inference to [their] judgment."

Distinct advantages of each scale.

Advantages of the Ottawa scale were associated with its ease of use, its succinct nature, and its more obvious translation to clinical practice. By comparison, advantages of the Chen scale were associated with its relationship to summative entrustment decisions.

Discussion

In this study, we compared the utility of 2 scales (Ottawa and Chen) for making ad hoc, formative decisions about medical students' clinical competence in the workplace by analyzing the ratings given by participants familiar with the Core EPAs to standardized learners in structured vignettes. Both scales performed well in differentiating between learners demonstrating less-developed and more-developed EPA skill levels; however, the Ottawa scale demonstrated relatively more desirable psychometric characteristics. While several previous studies have provided validity evidence for using the Ottawa scale with residents, [18,19,29–31] limited evidence supports using this scale in the UME setting. Across the UME-GME continuum, the validity evidence for the Chen scale is limited, [32] particularly when it is used to render formative, ad hoc entrustment decisions based on a single direct observation. A single published study analyzed the performance of the 2 scales in a UME setting at 1 institution with a small cohort of learners. [22] Thus, overall, the results of our study, albeit based on standardized learners, add to the existing validity evidence for both the Ottawa and Chen scales in discriminating between less-developed and more-developed EPA skill levels in medical students.

Validity evidence for the Ottawa and Chen supervisory scales

The Core EPA pilot has focused on developing WBAs that demonstrate validity and reliability and that directly reflect the daily work of physicians. However, the medical education community has not come to a consensus on the optimal rating scale for preceptors to use in the context of EPA-specific assessments for medical students. [6,14,33] In this section, we use Messick's framework [23] to summarize the existing validity evidence for the Ottawa and Chen scales. We then contextualize the results of our study within this existing evidence in the UME setting.

Content validity.

Rekman and colleagues recently reviewed existing "entrustability" scales. The authors suggest that several proposed scales, including the Ottawa scale, demonstrate alignment between the work of physicians and the construct of entrustment. [17] Further, content validity has been established for using the Ottawa scale in several surgical and internal medicine settings. [18,19,29,31] Chen and colleagues adapted the Chen scale based on supervisory scales in GME settings, extrapolating a developmental trajectory for the UME population. [20] However, no further content validity has been established for using the Chen scale in UME.

Cutrer and colleagues established content validity for both scales through local expert consensus generation. [22] In addition, the work of the Supervisory Taskforce of the Core EPA pilot has provided evidence for expert consensus. [21] The results of our study add to the content validity of both instruments.

Response process validity.

In the initial development of the Ottawa scale, focus group analysis found that raters appreciated the scale's alignment with summative decision making, which made it easy to use. [19] A subsequent study explored the perspectives of faculty and residents regarding the Ottawa scale compared with other, "traditional," summative, global rating scales. While participants appreciated the practical language of the Ottawa scale, they noted that context may create challenges (e.g., the scale may work better for observing procedures than for cognitive skills such as developing a differential diagnosis). [34] Cutrer and colleagues found that both the Ottawa and Chen scales assessed different attributes of learners, that a rater's preference for 1 scale was situationally determined (i.e., for tasks that naturally allow for independence such as documentation vs those that naturally require supervision such as handovers), and that there was substantial value in the narrative comments for the learners. [22]

Our study provides further insight into response process validity for the 2 entrustment scales. Content analysis of comments from participants assigned to the Ottawa scale revealed a consistent sentiment that the scale was easy to use and that it generally related to the intuitive perspectives of front-line faculty. The participants assigned to the Chen scale appreciated how the scale directly related to an ultimate entrustment decision about the level of supervision required and that it was forward thinking in nature. However, participants noted both scales worked well for particular EPAs but not others.

Internal structure validity.

Much of the validity evidence around the Ottawa scale has centered on internal structure evidence. [18,19,29,31] The original study demonstrated high reliability between raters and items. [19] Subsequent studies across specialties and diverse settings in GME have demonstrated similar results. [18,29,31] To the best of our knowledge, no other study has reported internal structure validity on the Chen scale. Cutrer and colleagues examined how the 2 scales aligned with one another but did not perform any psychometric analyses. [22]

To the best of our knowledge, our study is the first to assess and compare the internal structure validity evidence of both the Ottawa scale and Chen scale in a UME context. Overall, we observed similar reliability for both scales. Though this level of reliability may be inadequate for a high-stakes examination, it may be more acceptable for formative direct observation.

Compared with the Chen scale, the Ottawa scale had some advantageous attributes. First, the Ottawa scale demonstrated less variability across most EPAs for both less- and more-developed learners. This finding suggests that raters had more consensus on skill levels. Second, while the ICC was generally similar across EPAs and scales, we observed, for EPAs 2 and 5 (written notes), a notable, but not statistically significant, difference between ratings using the Ottawa scale and ratings using the Chen scale. We suspect this difference is secondary to the nature of anchors within the scales, particularly in the encounter note context. At the lower end of the Chen scale, anchors such as “1, Not allowed to practice” may not fit well with the nature of encounter note supervision. As our participants described, it is challenging to imagine a situation in which a faculty member would not *allow* a learner to attempt formulating a differential diagnosis or constructing an encounter note. In contrast, the entirety of the Ottawa scale may be more in line with faculty teaching and feedback practice because anchors relate to the degree of intervention required (e.g., “I had to,” “I had to talk them through”).

Finally, the G-study data are worth considering. While the variance attributed to the learner was similar (18.2% for Chen vs 18.5% for Ottawa), we observed a notable difference in encounter-based variance represented by *learner* × *EPA* (48.9% for Chen, 59.6% for Ottawa). This facet describes differences in learner performance for a given skill (i.e., EPA). A higher percentage attributed to this facet suggests greater sensitivity to identifying differences in the performance of learners on the unique EPAs. The implications of this difference are that the Ottawa scale may be more truly reflective of the learner’s performance for a given EPA and less reflective of factors unrelated to the learner or skill assessed.

Limitations

This study has several limitations. First, we chose 5 Core EPAs to study, so our findings cannot be generalized to the other 8 EPAs. The number and scope of vignettes in our study may have limited comparison between the 2 scales. Additionally, the structured setting of the study may not resemble the real-life practice of raters because they were not able to interact with learners. Further, by design, our study participants were experienced medical educators, most of whom had 10 or more years of clinical teaching experience and were familiar with the EPA framework. Less experienced raters may use the 2 scales differently than raters with a shared mental model. Finally, we did not consider the potential for bias by raters in terms of the learner's characteristics or those of the raters themselves.

Conclusions and Next Steps

Entrustment decisions in the UME context are ultimately binary; a learner is either preentrustable (not yet ready for indirect supervision) or entrustable (ready for indirect supervision). [12] Data from this study demonstrate that, in a highly structured environment, both the Ottawa scale and Chen scale performed well in discriminating skills of preentrustable and entrustable learners. Therefore, these findings represent a critical step forward in developing instruments with validity evidence to measure medical student progression toward entrustment in the Core EPAs. To further determine the value of these scales, next steps include implementation and analysis in the authentic workplace with faculty who do not have experience with the Core EPAs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

The authors would like to thank Sarah Russel, MD (a recent graduate of the University of Illinois College of Medicine), who served as the standardized medical student in the video-based vignettes, for her contribution to this study. They would also like to thank the Association of American Medical Colleges (AAMC) Core EPA pilot faculty members for participating in this study.

Funding/Support:

The authors each serve as members of the AAMC Core EPA pilot. Team members from each institution receive travel support to attend meetings with other pilot members and leaders from the AAMC. Dr. Santen has funding from Clinical and Translational Science Award No. UL1TR002649 from the National Center for Advancing Translational Sciences.

References

1. Lockyer J, Carraccio C, Chan MK, et al. ; ICBME Collaborators. Core principles of assessment in competency-based medical education. Med Teach. 2017;39:609–616. [PubMed: 28598746]
2. Bok HG, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: When theory meets practice. BMC Med Educ. 2013;13:123. [PubMed: 24020944]
3. ten Cate O, Snell L, Carraccio C. Medical competence: The interplay between individual ability and the health care environment. Med Teach. 2010;32:669–675. [PubMed: 20662579]
4. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32:676–682. [PubMed: 20662580]
5. Hawkins RE, Welcher CM, Holmboe ES, et al. Implementation of competency-based medical education: Are we addressing the concerns and challenges? Med Educ. 2015;49:1086–1102. [PubMed: 26494062]
6. Lomis K, Amiel JM, Ryan MS, et al. ; AAMC Core EPAs for Entering Residency Pilot Team. Implementing an entrustable professional activities framework in undergraduate medical education: Early lessons from the AAMC Core Entrustable Professional Activities for Entering Residency pilot. Acad Med. 2017;92:765–770. [PubMed: 28557937]

7. Lupi CS, Ownby AR, Jokela JA, et al. ; Association of American Medical Colleges Core Entrustable Professional Activities for Entering Residency Faculty Development Concept Group. Faculty development revisited: A systems-based view of stakeholder development to meet the demands of entrustable professional activity implementation. Acad Med. 2018;93:1472–1479. [PubMed: 29794524]

8. Elnicki DM, Aiyer MK, Cannarozzi ML, et al. An entrustable professional activity (EPA)-based framework to prepare fourth-year medical students for internal medicine careers. J Gen Intern Med. 2017;32:1255–1260. [PubMed: 28634908]

9. Lindeman BM, Sacks BC, Lipsett PA. Graduating students' and surgery program directors' views of the Association of American Medical Colleges Core Entrustable Professional Activities for Entering Residency: Where are the gaps? J Surg Educ. 2015;72:e184–e192. [PubMed: 26276302]

10. Pearlman RE, Pawelczak M, Yacht AC, Akbar S, Farina GA. Program director perceptions of proficiency in the core entrustable professional activities. J Grad Med Educ. 2017;9:588–592. [PubMed: 29075377]

11. Ryan MS, Lockeman KS, Feldman M, Dow A. The gap between current and ideal approaches to the core EPAs: A mixed methods study of recent medical school graduates. MedSciEduc. 2016;26:463–473.

12. Englander R, Flynn T, Call S, et al. Toward defining the foundation of the MD degree: Core entrustable professional activities for entering residency. Acad Med. 2016;91:1352–1358. [PubMed: 27097053]

13. Association of American Medical Colleges. Core EPAs pilot participants. http://www.aamc.org/initiatives/coreepas/pilotparticipants. Published 2019. Accessed August 29, 2019.

14. ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE guide no. 99. Med Teach. 2015;37:983–1002. [PubMed: 26172347]

15. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. Med Teach. 2007;29:855–871. [PubMed: 18158655]

16. Hurst YK, Prescott-Clements L. Optimising workplace-based assessment. Clin Teach. 2018;15:7–12. [PubMed: 29333757]

17. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: Outlining their usefulness for competency-based clinical assessment. Acad Med. 2016;91:186–190. [PubMed: 26630609]

18. Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W. A new instrument for assessing resident competence in surgical clinic: The Ottawa Clinic Assessment Tool. J Surg Educ. 2016;73:575–582. [PubMed: 27052202]

19. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–1407. [PubMed: 22914526]

20. Chen HC, van den Broek WE, ten Cate O. The case for use of entrustable professional activities in undergraduate medical education. Acad Med. 2015;90:431–436. [PubMed: 25470310]

21. Obeso V, Brown D, Aiyer M, et al. Core Entrustable Professional Activities for Entering Residency: Toolkits for the 13 Core EPAs. Washington, DC: Association of American Medical Colleges; 2017. http://www.aamc.org/what-we-do/mission-areas/medical-education/cbme/core-epas/publications. Accessed June 9, 2021.

22. Cutrer WB, Russell RG, Davidson M, Lomis KD. Assessing medical student performance of Entrustable Professional Activities: A mixed methods comparison of Co-Activity and Supervisory Scales. Med Teach. 2020;42:325–332. [PubMed: 31714166]

23. Messick S. Standards of validity and the validity of standards in performance assessment. Educ Meas. 1995;14:5–8.

24. Calaman S, Hepps JH, Bismilla Z, et al. ; I-PASS Study Education Executive Committee. The creation of standard-setting videos to support faculty observations of learner performance and entrustment decisions. Acad Med. 2016;91:204–209. [PubMed: 26266461]

25. Spector ND, Starner AJ, Allen AD, et al. I-PASS Handoff Curriculum: Core Resident Workshop. MedEdPORTAL. 2013. doi: 10.15766/mep_2374-8265.9311.

26. Kreiter CD, Zaidi NL, Park YS. Chapter 4: Generalizability theory. In: Yudkowky R, Park YS, Downing SM, eds. Assessment in Health Professions Education. 2nd ed. New York, NY: Routledge; 2020;51–69.

27. Krippendorff K. Content Analysis: An Introduction to Its Methodology. 4th ed. Thousand Oaks, CA: Sage Publications; 2018.

28. Cicchetti D. , Guidelines criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6:284–290.

29. MacEwan MJ, Dudek NL, Wood TJ, Gofton WT. Continued validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): Use in the simulated environment. Teach Learn Med. 2016;28:72–79. [PubMed: 26787087]

30. Saliken D, Dudek N, Wood TJ, MacEwan M, Gofton WT. Comparison of the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE) to a single-item performance score. Teach Learn Med. 2019;31:146–153. [PubMed: 30514128]

31. Halman S, Rekman J, Wood T, Baird A, Gofton W, Dudek N. Avoid reinventing the wheel: Implementation of the Ottawa Clinic Assessment Tool (OCAT) in Internal Medicine. BMC Med Educ. 2018; 18:218. [PubMed: 30236097]

32. Schumacher DJ, West DC, Schwartz A, et al. ; Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network General Pediatrics Entrustable Professional Activities Study Group. Longitudinal assessment of resident performance using entrustable professional activities. JAMA Netw Open. 2020;3:e1919316.

33. Brown DR, Warren JB, Hyderi A, et al. ; AAMC Core Entrustable Professional Activities for Entering Residency Entrustment Concept Group. Finding a path to entrustment in undergraduate medical education: A progress report from the AAMC Core Entrustable Professional Activities for Entering Residency Entrustment Concept Group. Acad Med. 2017;92:774–779. [PubMed: 28557941]

34. Dudek N, Gofton W, Rekman J, McDougall A. Faculty and resident perspectives on using entrustment anchors for workplace-based assessment. J Grad Med Educ. 2019;11:287–294. [PubMed: 31210859]

Table 1

Case Number, Learner Level, EPA Assessed, and Type and Source of Vignette for a Study Comparing the Performance of the Ottawa Scale vs the Chen Scale in Measuring the Performance of a Standardized Learner, 2019

Case #	Learner level	EPA(s) assessed	Type	Source
1	Preentrustable	2, 6	Video	Author developed
2	Entrustable	2, 5	Written encounter note	Author developed
3	Entrustable	2, 6	Video	Author developed
4	Preentrustable	1	Video	Author developed
5	Preentrustable	8	Video	Spector et al[25]
6	Entrustable	1	Video	Author developed
7	Preentrustable	2, 5	Written encounter note	Author developed
8	Entrustable	8	Video	Spector et al[25]

Abbreviation: Core EPA, Core Entrustable Professional Activity for Entering Residency.

Table 2

Summary of Demographic Characteristics of 50 Faculty Participants Assessing Case-Based, Video-Recorded Vignettes of a Standardized Learning Executing Core EPAs, 2019

Characteristic	Group using the Ottawa scale, no. (% of 22)	Group using the Chen scale, no. (% of 28)
Institution		
Columbia	1 (4.5)	2 (7.1)
Florida International University Herbert Wertheim College of Medicine	1 (4.5)	2 (7.1)
Michigan State University College of Human Medicine	1 (4.5)	2 (7.1)
New York University School of Medicine	0	1 (3.6)
Oregon Health & Science University School of Medicine	4 (18.2)	3 (10.7)
University of Texas Science Center at Houston	2 (9.1)	1 (3.6)
University of Illinois College of Medicine	8 (36.4)	6 (21.4)
Vanderbilt	2 (9.1)	4 (14.3)
Virginia Commonwealth University School of Medicine	2 (9.1)	5 (17.9)
Yale	1 (4.5)	1 (3.6)
None listed	0	1 (3.6)
Specialty		
Anesthesiology	0	1 (3.6)
Emergency medicine	1 (4.5)	0
Family medicine	1 (4.5)	1 (3.6)
Internal medicine	13 (59.1)	16 (57.1)
Internal medicine/pediatrics	1 (4.5)	1 (3.6)
Neurology	0	1 (3.6)
Pediatrics	4 (18.1)	5 (17.9)
Psychiatry	1 (4.5)	1 (3.6)
Surgery	1 (4.5)	2 (7.1)
Gender		
Female	14 (63.6)	12 (42.9)
Male	8 (36.4)	15 (53.6)
Not listed	0	1 (3.6)
Any prior experience with Core EPA-based WBAs?		
Yes	13 (59.1)	18 (64.3)

Characteristic	Group using the Ottawa scale, no. (% of 22)	Group using the Chen scale, no. (% of 28)
No	9(41.0)	10 (35.7)
Years working with/assessing learners		
< 5 years	2 (9.1)	5 (17.9)
> 5–10 years	8 (36.4)	9 (32.1)
> 10 years	12 (54.5)	14 (50.0)

Abbreviations: Core EPA, Core Entrustable Professional Activity for Entering Residency; WBA, workplace-based assessment.

Table 3

Descriptive Statistics Comparing Scores Using the Ottawa Scale vs the Chen Scale for Rating the Performance of Less- and More-Developed Learners by Core EPA, 2019[a]

			Less-developed learner		More-developed learner		
Core EPA #: Description	**Context of observation**	**Scale**	**Mean (SD)**	**Range**	**Mean (SD)**	**Range**	***P* value[b]**
1: History and physical	History and physical	Ottawa	2.48 (0.93)	1–5	4.62 (0.50)	4–5	< .001
		Chen	2.67 (0.96)	1–5	4.89 (0.32)	4–5	< .001
2: Differential diagnosis	Oral case presentation	Ottawa	2.13 (0.76)	1–4	4.24 (0.62)	3–5	< .001
		Chen	2.82 (0.90)	1–4	4.64 (0.56)	3–5	< .001
2: Differential diagnosis	Written encounter note	Ottawa	1.64 (0.58)	1–3	3.77 (0.69)	3–5	< .001
		Chen	2.63 (0.93)	1–5	4.07 (0.86)	2–5	< .001
5: Encounter note	Written encounter note	Ottawa	1.86 (0.57)	1–3	3.76 (0.89)	2–5	< .001
		Chen	2.77 (0.99)	1–5	4.32 (0.77)	2–5	< .001
6: Oral presentation	Oral case presentation	Ottawa	2.68 (0.65)	1–4	4.24 (0.83)	3–5	< .001
		Chen	3.07 (0.94)	1–5	4.64 (0.56)	3–5	< .001
8: Handover	Patient handover	Ottawa	2.00 (0.46)	1–3	4.47 (0.61)	3–5	< .001
		Chen	2.74 (0.81)	1–5	4.65 (0.63)	3–5	< .001

Abbreviations: Core EPA, Core Entrustable Professional Activity for Entering Residency; SD, standard deviation.

[a] Fifty faculty members from 10 institutions produced, collectively, 579 Core EPA assessments.

[b] The *P* values are based on *t* tests comparing differences in mean ratings between less- and more-developed learners.

Table 4

Interrater Reliability: ICC and 95% CIs Comparing Ottawa vs Chen Scale for Rating the Performance of a Simulated Learner by Core EPA, 2019[a]

		Group using the Ottawa scale[b]		Group using the Chen scale[b]	
Core EPA #: Description	**Context of observation**	**ICC**	**95% CI**	**ICC**	**95% CI**
1: History and physical	History and physical	0.79	0.40, 1.00	0.83	0.45, 1.00
2: Differential diagnosis	Oral case presentation	0.82	0.45, 1.00	0.74	0.34, 1.00
2: Differential diagnosis	Written note	0.84	0.48, 1.00	0.54	0.17, 1.00
5 Encounter note	Written note	0.75	0.34, 1.00	0.58	0.18, 1.00
6: Oral presentation	Oral case presentation	0.68	0.26, 1.00	0.67	0.27, 1.00
8: Handover	Patient handover	0.91	0.64, 1.00	0.75	0.35, 1.00
Summary		0.80	0.43, 1.00	0.69	0.29, 1.00

Abbreviations: ICC, intraclass correlation; CI, confidence interval; Core EPA, Core Entrustable Professional Activity for Entering Residency.

[a] Fifty faculty members from 10 institutions produced, collectively, 579 Core EPA assessments.

[b] Internal-consistency reliability (Cronbach's alpha) for the Chen and Ottawa scales are 0.70 and 0.67, respectively.

Chart 1

Results of a Generalizability Study Comparing the Performance of the Ottawa Scale vs the Chen Scale in Measuring the Performance of a Standardized Learner, by Core EPA, 2019[a]

		Ottawa scale		**Chen scale**	
Type	**Effect**	**VC**	**% VC**	**VC**	**% VC**
VC	Learner	0.417	18.5	0.379	18.2
	Rater	0.052	2.3	0.127	6.1
	EPA	0.000	0.0	0.000	0.0
	Learner x rater	0.040	1.8	0.050	2.4
	Learner x EPA	1.347	59.6	1.021	48.9
	Rater x EPA	0.000	0.0	0.000	0.0
	Residual	0.403	17.8	0.511	24.5
Reliability	G-coefficient	0.645		0.684	
	ϕ-coefficient	0.643		0.677	
SEM	Relative SEM	0.479		0.418	
	Absolute SEM	0.481		0.425	

Abbreviations: Core EPA, Core Entrustable Professional Activity for Entering Residency; VC, variance component; % VC, percentage of variance due to component; SEM, standard error of measurement.

[a] Data are based on 579 Core EPA assessments produced by 50 faculty participants from 10 institutions.