Exploring the Association Between USMLE Scores and ACGME Milestone Ratings: A Validity Study Using National Data From Emergency Medicine

Stanley J. Hamstra, PhD, Monica M. Cuddy, MA, Daniel Jurich, PhD, Kenji Yamazaki, PhD, John Burkhardt, MD, PhD, Eric S. Holmboe, MD, Michael A. Barone, MD, MPH, and Sally A. Santen, MD, PhD

Abstract

Purpose

The United States Medical Licensing Examination (USMLE) sequence and the Accreditation Council for Graduate Medical Education (ACGME) milestones represent 2 major components along the continuum of assessment from undergraduate through graduate medical education. This study examines associations between USMLE Step 1 and Step 2 Clinical Knowledge (CK) scores and ACGME emergency medicine (EM) milestone ratings.

Method

In February 2019, subject matter experts (SMEs) provided judgments of expected associations for each combination of Step examination and EM subcompetency. The resulting sets of subcompetencies with expected strong and weak associations were selected for convergent and discriminant validity analysis, respectively. National-level data for 2013–2018 were provided; the final sample included 6,618 EM residents from 158 training programs. Empirical bivariate correlations between milestone ratings and Step scores were calculated, then those correlations were compared with the SMEs' judgments. Multilevel regression analyses were conducted on the selected subcompetencies, in which milestone ratings were the dependent variable, and Step 1 score, Step 2 CK score, and cohort year were independent variables.

Results

Regression results showed small but statistically significant positive relationships between Step 2 CK score and the subcompetencies (regression coefficients ranged from 0.02 [95% confidence interval (CI), 0.01–0.03] to 0.12 [95% CI, 0.11–0.13]; all $P < .05$), with the degree of association matching the SMEs' judgments for 7 of the 9 selected subcompetencies. For example, a 1 standard deviation increase in Step 2 CK score predicted a 0.12 increase in MK-01 milestone rating, when controlling for Step 1. Step 1 score showed a small statistically significant effect with only the MK-01 subcompetency (regression coefficient = 0.06 [95% CI, 0.05–0.07], $P < .05$).

Conclusions

These results provide incremental validity evidence in support of Step 1 and Step 2 CK score and EM milestone rating uses.

Physicians pass through a *continuum of assessment* as they progress from undergraduate through graduate medical education. This continuum of assessment provides documentation of evolving knowledge, skills, and abilities. One useful framework for describing this growth is Miller's pyramid, which includes the progressive stages of knows, knows how, shows how, and does (see Table 1).[1,2] Multiple-choice question examinations are useful for measuring medical knowledge (knows) and its application (knows how), while standardized performance-based assessments may be best for measuring performance in simulated clinical environments (shows how), and direct observation in clinical settings may be particularly well suited for measuring work performance (does). Reflecting different phases in the continuum of assessment, the United States Medical Licensing Examination (USMLE) sequence and the Accreditation Council for Graduate Medical Education (ACGME) milestones are designed for evaluating competencies deemed essential for safe and effective practice.

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

Correspondence should be addressed to Stanley J. Hamstra, Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, 2075 Bayview Ave., Room A3-31, Toronto, ON M4N 3M5, Canada; telephone: (437) 881-6191; email: stan.hamstra@utoronto.ca; Twitter: @stanhamstra.

Acad Med. 2021;96:1324–1331.
First published online June 15, 2021
doi: 10.1097/ACM.0000000000004207

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/B134 and http://links.lww.com/ACADMED/B135.

United States Medical Licensing Examination

To obtain a license to practice medicine in the United States, graduates of medical schools must pass the USMLE sequence, consisting of 3 steps, each of which assesses distinct knowledge domains and skill sets. USMLE Step scores are intended to reflect a progression of knowledge accumulation and skills acquisition that build on each other and link to different developmental phases. One way in which the USMLE sequence measures the accumulation and application of knowledge (or knows and knows how) is with multiple-choice questions presented as clinical vignettes, some of which include multimedia elements to better approximate clinical experiences.

In this study, we focus on USMLE Step 1 and Step 2 Clinical Knowledge (CK). Step 1 is designed to assess an individual's ability to understand and apply basic science concepts important to the practice of medicine. Step 2 CK targets an individual's ability to apply medical knowledge, clinical science, and clinical skills necessary for patient care under supervision.

Table 1
Methods of Assessment Along the Continuum of Medical Education Suggested by Miller's Pyramid[1,a]

Miller's level	Focus of the assessment	Methods
Does	Performance in the clinical workplace	Tools for direct observation (mini-CEX, DOPS), case-based discussions, multisource feedback
Shows how	Standardized psychometric assessment of performance associated with real clinical encounters	OSCEs, simulations with artifacts (e.g., virtual reality tools, procedural task trainers), mannequin-based cases or standardized patients
Knows how	Ability to apply knowledge to clinical problems	Written and online tests with a variety of problem-solving approaches, oral tests
Knows	Knowledge of the facts and processes relevant to clinical problems	Written and online tests

Abbreviations: mini-CEX, mini-clinical evaluation exercise; DOPS, direct observation of procedural skills; OSCEs, objective structured clinical examinations.
[a]Adapted from ten Cate O, Carraccio C, Damodaran A, et al. Entrustment decision making: Extending Miller's Pyramid. Acad Med. 2021;96:199–204.[2]

ACGME milestones

The ACGME milestones are used by residency programs to monitor the progression of competence in the clinical practice environment. The major core competencies of the milestones are patient care (PC), medical knowledge (MK), professionalism (PR), interpersonal and communication skills (ICS), systems-based practice (SBP), and practice-based learning and improvement, each of which consists of several subcompetencies. The milestones are designed to evaluate whether what has been learned in controlled settings (or shows how) can be effectively translated to actual practice (or does). Milestone ratings are generated by the residency program's clinical competency committee and reported to the ACGME every 6 months during residency training. Milestones and subcompetencies are formulated to be specialty specific. Within each subcompetency, narrative anchors are provided for monitoring progression across 5 levels (see Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/B134).

Collecting validity evidence across the continuum of assessment

Valid score interpretations depend on the degree to which USMLE scores and ACGME milestone ratings can be translated into inferences about competency and, ultimately, into entrustment decisions. Both the USMLE Step examinations and the ACGME milestones require ongoing evaluation of validity evidence to support their uses and interpretations of results.[3–6] Evidence for content validity can be obtained from subject matter experts (SMEs), whose detailed knowledge of both the Step exams and the milestones offers a unique comparison of content of 2 independently constructed assessment systems that have been deemed important to the profession. Evidence based on relations with other variables can be obtained by correlating Step scores with milestone ratings. Convergent validity evidence is demonstrated when 2 assessments intended to measure similar constructs (e.g., MK milestone ratings and Step 1 scores) yield performance outcomes that are positively correlated. Conversely, discriminant validity evidence is demonstrated when assessments intended to measure dissimilar constructs (e.g., ICS milestone ratings and Step 1 scores) yield weak or no empirical correlations. Measuring both convergent and discriminant validity evidence *simultaneously* tests the relative strength of predicted relationships among different dimensions or subscales and is, therefore, less likely to be due to chance than overall post hoc correlations.[3,6,7]

Rationale and purpose of the research

Given that USMLE scores and ACGME milestone ratings (1) derive from independent assessments of the same learners and (2) form a longitudinal dataset that spans from undergraduate to graduate medical education when merged, they afford a unique opportunity to study the validity of both systems simultaneously. Although their purposes differ, there is some overlap in the constructs the 2 assessment systems measure. At the same time, they focus on different parts of the continuum of assessment outlined by Miller's pyramid.[1,2] Thus, certain elements of each assessment system would be expected to correlate positively, while others would be expected to show very weak or no correlation.

The purpose of the current study was to examine the associations between USMLE Step 1 and Step 2 CK scores collected during medical school and ACGME milestone ratings in emergency medicine (EM) collected at the end of postgraduate year 1 (PGY-1). We focus on Step 1 and Step 2 CK because they typically are taken before entry into graduate medical education. In addition, we chose to focus on milestone ratings in EM because there has been preliminary validity work on their internal structure and relations with other variables.[8–10] The complete list of the 23 EM subcompetencies and a brief description of their content are given in Supplemental Digital Appendix 2 (at http://links.lww.com/ACADMED/B134).

Method

Use of SMEs

We employed a group of SMEs with content knowledge of both the USMLE Step examinations and EM milestones to help focus the statistical analysis using a convergent and discriminant validity evidence design. This approach was taken to avoid an atheoretical examination of all possible bivariate correlations between Step 1 and Step 2 CK scores and milestone ratings from all 23 EM subcompetencies. In

analytic terms, this approach effectively lowered the probability of false positives (type I errors) by reducing the focus of the analysis to a subset of EM subcompetencies with the most or least expected congruence with Step scores.

SMEs included clinician-educators who served on the National Board of Medical Examiners (NBME) Emergency Medicine Advanced Clinical Examination Task Force. These task force members are responsible for the development of the NBME Emergency Medicine Advanced Clinical Examination and include senior educators with experience in both undergraduate (curricular deans and clerkship directors) and graduate medical education (residency program leadership).[11] Ten task force members were invited to participate; 7 completed the voluntary exercise.

In February 2019, SMEs were asked to judge the expected strength of association for each combination of Step examination and EM subcompetency, using a scale of 0–3, where 0 = no association and 3 = strong correlation. The instructions provided to the SMEs are detailed in Supplemental Digital Appendix 3 (at http://links.lww.com/ACADMED/B135). Mean SME ratings were calculated for each EM subcompetency, resulting in a rank-ordered list of EM subcompetencies according to the strength of expected associations with Step scores. The resulting set of EM subcompetencies with expected strong associations with the Step examinations included MK-01, PC-05, and PC-04 and the resulting set with expected weak associations included PC-06, PR-01, PC-08, PC-09, SBP-02, and ICS-01. We selected these 9 subcompetencies for further analysis. The remaining subcompetencies were not the primary focus of this analysis; the full rank-ordered list of subcompetencies is provided in Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/B135).

Study sample and data

National-level USMLE score and EM milestone data were provided by the NBME and ACGME, respectively. The ACGME provided PGY-1 milestone ratings and residency program data for EM from 2013 to 2018, yielding an initial sample of data from 9,547 residents. Residents who graduated from medical schools outside of the United States (n = 1,757; 18.4%) and those with osteopathic degrees (DOs; n = 1,031; 13.2% of the remaining residents in the U.S. sample) were excluded due to differences in training and pathways to licensure. The NBME provided USMLE Step 1 and Step 2 CK scores and medical school characteristic variables (for matching purposes). Combining the datasets yielded a 99.9% match rate and resulted in a sample of 6,728 EM residents from 176 training programs. Matched residents who took Step 1 or Step 2 CK before 2010 (n = 110; 1.6%) were excluded to minimize the amount of time between the completion of the Step examinations and the end of the first year of residency training. The final sample used for analysis included 6,618 EM residents from 158 training programs. This is the first time that national datasets from the NBME and ACGME have been merged.

Expected associations

Empirical bivariate correlations were calculated to examine the relationships between Step 1 and Step 2 CK scores and milestone ratings for the selected EM subcompetencies. The strengths of these correlations were then compared with the SME's judgments to determine whether data from both assessment systems behaved as expected with respect to convergent and discriminant validity evidence.

Multilevel regression

To address known variation in milestone ratings at the program level,[10,12] multilevel regression techniques were used.[13,14] An intercept-only model was estimated to calculate intraclass correlation coefficients for each of the selected EM subcompetencies. This model allowed for an estimate of the amount of variance in milestone ratings due to differences between residency programs. Intraclass correlation coefficients ranged from 0.27 to 0.41, indicating large proportions of variance in milestone ratings at the program level. This finding justified the use of multilevel regression techniques.

Next, for each selected EM subcompetency, a model was estimated to predict milestone ratings (dependent variable) based on Step 1 score, Step 2 CK score, and cohort year (independent variables). Cohort year was included as a categorical variable to control for differences in cohorts across the time frame of the study based on previous research.[10] In all models, cohort year was treated as a random effect and allowed to vary across programs. Preliminary analyses indicated that the relationships between Step 1 and Step 2 CK scores and EM milestone ratings did not vary considerably across programs. As such, the effects of Step 1 and Step 2 CK scores were treated as fixed effects, where the same association was estimated for all programs. Step scores were standardized to a mean of 0 and a standard deviation (SD) of 1 so that regression coefficients could be interpreted as the average change in milestone rating for every 1 SD change in Step score. For each model, the practical and statistical effects of the regression coefficients were (1) evaluated to determine whether they aligned with expectations and (2) compared to determine the extent to which each of the Step scores (i.e., Step 1 and Step 2 CK scores) contributed unique information to understanding variation in EM milestone ratings.

This study was approved by the institutional review board of the American Institutes for Research (AIR EX00490) on August 11, 2019.

Results

Study sample and summary statistics

The majority of the 6,618 EM residents in the study sample were male (n = 4,301; 65.0%), and the average age was 29 years (range, 24–55) at the end of PGY-1. Table 2 provides descriptive statistics for the Step scores and EM milestone ratings used in this study.

Expected associations

Table 3 shows mean SME judgments about the expected strength of associations between Step scores and milestone ratings for the selected EM subcompetencies. It also presents empirical bivariate correlations. In some instances, the SMEs' expectations aligned with the empirical results when rank ordering the magnitude of the bivariate correlations within Step examination. For example, Step 1 scores showed the strongest bivariate correlations with ratings for the MK-01 subcompetency. Ratings for the PC subcompetencies (PC-04, PC-05, PC-06, PC-08, and PC-09) showed similar

Table 2

Mean (Standard Deviation) of National EM Milestone Ratings at the End of PGY-1 for Selected Subcompetencies and USMLE Step Scores for Study Sample[a]

Variable	2013 (n = 1,034)	2014 (n = 1,046)	2015 (n = 1,040)	2016 (n = 1,101)	2017 (n = 1,193)	2018 (n = 1,204)	Total (N = 6,618)
MK-01	2.11 (0.67)	2.02 (0.57)	2.01 (0.60)	2.03 (0.59)	2.05 (0.57)	2.01 (0.55)	2.04 (0.59)
PC-05	2.15 (0.51)	2.09 (0.52)	2.02 (0.50)	2.00 (0.47)	2.01 (0.43)	2.02 (0.44)	2.05 (0.48)
PC-04	2.28 (0.52)	2.26 (0.54)	2.14 (0.48)	2.12 (0.46)	2.12 (0.44)	2.11 (0.45)	2.17 (0.49)
PC-06	2.31 (0.51)	2.25 (0.48)	2.15 (0.48)	2.12 (0.49)	2.11 (0.45)	2.11 (0.46)	2.17 (0.48)
PR-01	2.40 (0.53)	2.33 (0.52)	2.28 (0.55)	2.26 (0.51)	2.22 (0.49)	2.21 (0.51)	2.28 (0.52)
PC-08	2.24 (0.52)	2.18 (0.47)	2.10 (0.48)	2.12 (0.49)	2.09 (0.45)	2.08 (0.45)	2.13 (0.48)
PC-09	2.13 (0.48)	2.10 (0.52)	2.09 (0.52)	2.06 (0.51)	2.03 (0.44)	2.02 (0.46)	2.07 (0.49)
SBP-02	2.21 (0.49)	2.15 (0.47)	2.10 (0.45)	2.07 (0.44)	2.07 (0.39)	2.08 (0.44)	2.11 (0.45)
ICS-01	2.38 (0.56)	2.28 (0.53)	2.20 (0.53)	2.21 (0.52)	2.19 (0.47)	2.18 (0.48)	2.24 (0.52)
Step 1	226 (18)	228 (18)	229 (16)	232 (16)	231 (15)	231 (16)	229 (17)
Step 2 CK	241 (16)	241 (16)	242 (15)	244 (14)	244 (14)	245 (14)	243 (15)

Abbreviations: EM, emergency medicine; PGY-1, postgraduate year 1; USMLE, United States Medical Licensing Examination; MK, medical knowledge; PC, patient care; PR, professionalism; SBP, systems-based practice; ICS, interpersonal and communication skills; CK, Clinical Knowledge.

[a]This study included 6,618 PGY-1 EM residents from 158 U.S. residency programs from 2013 to 2018. Milestone ratings were on a 5-point scale, ranging from 0 to 5 (where 0 = level expected of novice resident and 5 = aspirational level beyond that expected at time of graduation), in increments of 0.5. The subcompetencies shown here do not represent the full complement of EM subcompetencies but were selected for analysis for the purposes of this study (see main text for more information). Step 1 and Step 2 CK scores were out of a total possible range of 1 to 300.

correlations with Step 1 scores regardless of the strength of association expected by the SMEs (strong or weak). PR-01 and ICS-01 showed the weakest correlations with Step 1 scores, although the SMEs did not differentiate among the lowest expected associations with Step 1 scores (0.29 for all subcompetencies). Like Step 1, Step 2 CK scores showed the strongest correlations for the MK-01 subcompetency ratings. The subcompetencies with the lowest correlations for Step 2 CK were among the group of subcompetencies that the SMEs expected to have the lowest associations with Step 2 CK scores (PR-01 and ICS-01). For both Step examinations, PC-05 and PC-04 ratings showed considerably weaker correlations than expected by the SMEs. Among the group of subcompetencies that the SMEs expected to have strong associations, the expected associations were greater for Step 2 CK than for Step 1, and in all cases, the correlations were higher for Step 2 CK scores than for Step 1 scores.

Table 3

Mean SME Ratings (Rank Ordered From Highest to Lowest) and Bivariate Correlations Between EM Milestone Subcompetency Ratings and USMLE Scores[a]

	Expected strength of association, mean (range)[b]		Empirical bivariate correlations	
Subcompetency	Step 1	Step 2 CK	Step 1	Step 2 CK
Expected strong associations (convergent validity)				
MK-01	2.14 (1–3)	2.43 (2–3)	0.23	0.26
PC-05	1.43 (0–3)	2.00 (1–3)	0.08	0.12
PC-04	1.00 (0–2)	1.57 (1–3)	0.11	0.15
Expected weak associations (discriminant validity)				
PC-06	0.29 (0–1)	0.43 (0–1)	0.08	0.11
PR-01	0.29 (0–1)	0.43 (0–1)	0.03	0.04
PC-08	0.29 (0–1)	0.29 (0–1)	0.10	0.14
PC-09	0.29 (0–1)	0.29 (0–1)	0.08	0.09
SBP-02	0.29 (0–1)	0.29 (0–1)	0.08	0.09
ICS-01	0.29 (0–1)	0.29 (0–1)	0.05	0.07

Abbreviations: SMEs, subject matter experts; EM, emergency medicine; USMLE, United States Medical Licensing Examination; CK, Clinical Knowledge; MK, medical knowledge; PC, patient care; PR, professionalism; SBP, systems-based practice; ICS, interpersonal and communication skills; PGY-1, postgraduate year 1.

[a]This study included 6,618 PGY-1 EM residents from 158 U.S. residency programs from 2013 to 2018. The subcompetencies shown here do not represent the full complement of EM subcompetencies but were selected for analysis for the purposes of this study (see main text for more information).

[b]Out of a possible range of 0 to 3, where 0 = no association and 3 = strong correlation.

To provide additional context, Figure 1 displays the mean Step 2 CK scores and 95% confidence intervals (CIs) for residents at different milestone rating levels for MK-01 and PR-01. For MK-01, residents with higher Step 2 CK scores had, on average, higher MK-01 ratings at the end of PGY-1, especially over the range of ratings from 1.5 to 3.0. For PR-01, average Step 2 CK scores differed negligibly by rating level.

Multilevel regression estimates

Table 4 provides the results of the multilevel regression analyses, showing the slope coefficients associated with Step scores for each of the selected EM subcompetencies. Each coefficient reflects the predicted increase in milestone rating for a 1 SD increase in the corresponding Step score, after controlling for performance on the other Step examination. For example,

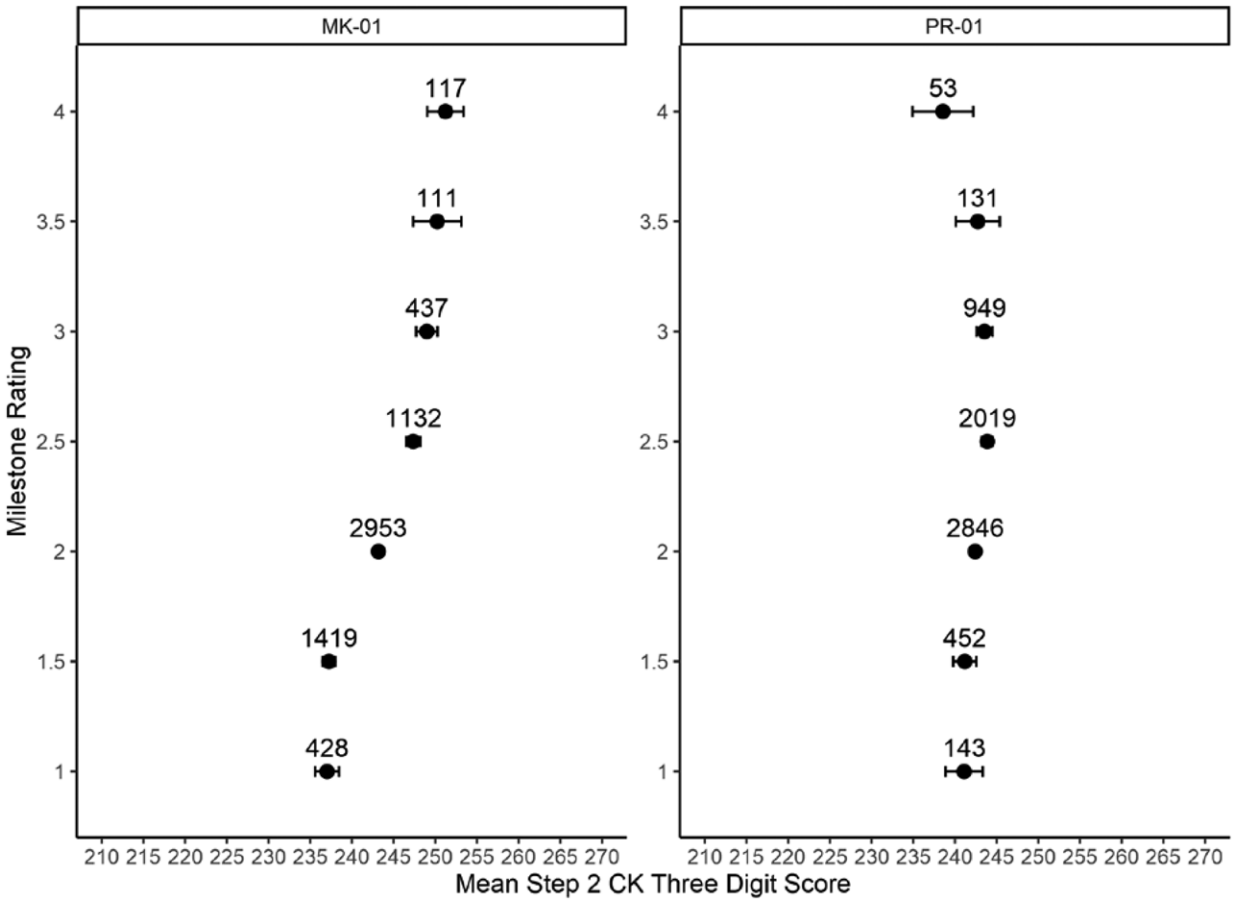

Figure 1 Mean Step 2 CK scores by milestone rating level for the MK-01 and PR-01 EM subcompetencies. From a study of 6,618 PGY-1 EM residents from 158 U.S. residency programs from 2013 to 2018. Error bars reflect 95% confidence interval for the mean. Note only milestone ratings with more than 30 residents are shown. Error bars may be smaller than symbols. Milestone ratings ranging from 0 to 5 (where 0 = level expected of novice resident and 5 = aspirational level beyond that expected at time of graduation) in increments of 0.5. Step 2 CK scores were out of a total possible range of 1 to 300. Abbreviations: CK, Clinical Knowledge; MK, medical knowledge; PR, professionalism; EM, emergency medicine; PGY-1, postgraduate year 1.

the Step 2 CK coefficient for the MK-01 subcompetency indicates that a 1 SD increase in Step 2 CK score predicted a 0.12 increase in milestone rating, controlling for Step 1.

There was a small but statistically significant positive relationship between Step 2 CK score and milestone ratings for all selected EM subcompetencies (regression coefficients ranged from 0.02 [95% CI, 0.01–0.03] to 0.12 [95% CI, 0.11–0.13]; all $P < .05$), with the degree of association matching the SMEs' judgments for 7 of the 9 selected subcompetencies, after controlling for the other variables in the model (i.e., cohort year and Step 1 score). Step 1 score yielded a small statistically significant effect with only the MK-01 subcompetency (regression coefficient = 0.06 [95% CI, 0.05–0.07], $P < .05$) but showed nonsignificance with ratings from the other 8 subcompetencies, after controlling for cohort year and Step 2 CK score.

Like the bivariate correlations, the rank order of the regression coefficients aligned in some cases with the SMEs' expected associations. For example, the MK-01 subcompetency resulted in the largest regression coefficients for both Step 1 and Step 2 CK, while the SBP-02, PR-01, and ICS-01 subcompetencies were among the group that yielded the lowest regression coefficients. Although the rank order of the coefficients was somewhat consistent with expectations, as shown in Table 4, the magnitudes of the Step 1 and Step 2 CK coefficients were generally small.

Discussion

This study presents a national-level analysis of the association between USMLE Step 1 and Step 2 CK scores and ACGME EM milestone ratings. Using a unique dataset, it follows the same learners as they progress from undergraduate to graduate medical education and employs a multifaceted analytic approach. First, SMEs provided judgments about the extent to which they thought Step 1 and Step 2 CK scores would relate to milestone ratings for all 23 EM subcompetencies. The 3 subcompetencies with the highest and the 6 with the lowest expected associations were then selected for subsequent analysis. Overall, this work aligns with the commitment made by the Invitational Conference on USMLE Scoring to study correlations between USMLE performance and measures of residency performance and clinical practice.[15] It also adds to the existing body of validity evidence for USMLE scores[16–19] and ACGME milestone ratings.[8–10]

In some cases, the judgments of the SMEs appear consistent with the

Table 4

USMLE Step Examination Slope Coefficients From Final Multilevel Regression Models by EM Subcompetency (Rank Ordered From Strongest to Weakest Based on SMEs' Expected Strength of Association)[a]

Subcompetency	Step 1, coefficient (95% CI)	Step 2 CK, coefficient (95% CI)
Expected strong associations (convergent validity)		
MK-01	0.06[b] (0.05 to 0.07)	0.12[b] (0.11 to 0.13)
PC-05	0.00 (–0.01 to 0.01)	0.04[b] (0.03 to 0.05)
PC-04	0.00 (–0.01 to 0.01)	0.06[b] (0.05 to 0.07)
Expected weak associations (discriminant validity)		
PC-06	–0.01 (–0.02 to 0.01)	0.05[b] (0.04 to 0.06)
PR-01	–0.01 (–0.02 to 0.01)	0.02[b] (0.01 to 0.03)
PC-08	0.00 (–0.01 to 0.01)	0.06[b] (0.05 to 0.07)
PC-09	0.00 (–0.01 to 0.01)	0.03[b] (0.02 to 0.05)
SBP-02	0.00 (–0.01 to 0.01)	0.02[b] (0.01 to 0.03)
ICS-01	–0.01 (–0.02 to 0.01)	0.02[b] (0.01 to 0.03)

Abbreviations: USMLE, United States Medical Licensing Examination; EM, emergency medicine; SMEs, subject matter experts; CI, confidence interval; CK, Clinical Knowledge; MK, medical knowledge; PC, patient care; PR, professionalism; SBP, systems-based practice; ICS, interpersonal and communication skills; PGY-1, postgraduate year 1.

[a]This study included 6,618 PGY-1 EM residents from 158 U.S. residency programs from 2013 to 2018. The subcompetencies shown here do not represent the full complement of EM subcompetencies but were selected for analysis for the purposes of this study (see main text for more information).

[b]Statistically significant at $P < .05$.

rank ordering of empirical bivariate correlations and regression coefficients. These patterns suggest that Step 1 and Step 2 CK scores relate in certain ways to subsequent performance in medical practice as measured by EM milestone ratings. In this sense, they may reflect a learner's movement along a continuum of learning and assessment from knows and knows how to shows how and does within the framework of Miller's pyramid.[1,2] Furthermore, the SMEs' expectations tended to mirror the ranked bivariate correlations within each Step examination at both the high and low ends of the scale, providing convergent and discriminant validity evidence for interpretations of Step 1 and Step 2 CK scores and EM milestone ratings. From a content perspective, these results contribute to validity evidence. For example, both Step examinations and several milestone subcompetencies are designed to capture a learner's medical knowledge and, in turn, the highest correlations were observed for MK-01. While certain milestone subcompetencies reflect professionalism and interpersonal and communication skills, these are not the main focus of Step 1 and Step 2 CK; thus, they showed lower correlations with PR-01 and ICS-01.

Though some of the results are statistically significant, practically speaking they are small. And given the mixed results concerning PC-04 and PC-05, inferences should be made tentatively. While a case can be made that PC-04 behaved as expected, yielding the second highest correlation of all subcompetencies, the effect was slight, and for PC-05, the effect was no stronger than those for any of the subcompetencies with expected weak associations. This may be due to restrictions of range at both the high and low ends of the milestone rating scale. Indeed, 1 reason the effect for MK-01 (both for correlations and regressions) was considerably stronger than the PC subcompetencies may be suggested from Table 2. There, we see that the SDs are higher for MK-01 than any of the PC subcompetencies, suggesting that the narrow range of variance in the PC subcompetencies may have contributed to the lack of empirical correlations.

Despite these caveats, for both correlations and regressions, the Step 2 CK effect is larger than the Step 1 effect, possibly suggesting that performance on Step 2 CK may provide useful information for understanding performance in residency training in ways that performance on Step 1 may not. In addition, the effects for Step 2 CK are consistent with evidence for content validity when both examination and milestone content are considered. For example, Step 2 CK includes more content related to the diagnosis and management of illness and, thus, may be able to account for additional variation (above and beyond Step 1) in milestone ratings for subcompetencies that highlight certain aspects of patient care.

The delay between the completion of Step 1 and Step 2 CK and the end of PGY-1 training may have impacted learner competence in varying ways, thus, lessening the potential for observing a strong correlation with milestone ratings. This may be particularly true for Step 1, which typically is taken earlier than Step 2 CK. In addition, the 2 assessment systems involve different data collection formats: performance on Step 1 and Step 2 CK is based on standardized scores from high-stakes multiple-choice question examinations, while milestone ratings are determined by aggregating impressions during direct observations in the context of clinical practice over several occasions. Further research into these areas might be conducted via focused qualitative interviews that could shed new light on the USMLE sequence and the ACGME EM milestones, including, for example, exactly how the constructs measured by Step 2 CK align with specific subcompetencies. Lastly, much of the variation in milestone ratings is likely due to challenges with response process validity, involving collection and interpretation of assessment data by clinical competency committees.[20,21] Current work on an overhaul of the milestone system (Milestones 2.0) represents a response to these challenges.[22,23]

One limitation of this study is that the descriptor language for the MK-01 milestone levels contains specific reference to the USMLE Step 1 and Step 2 examinations. While program directors and clinical competency committees are free to determine how they rate competency in medical knowledge, it is possible that many used prior awareness of examination scores in assigning milestone ratings for this subcompetency. To address this more directly, future studies could examine in more detail

associations between specialty in-training examination scores and milestone ratings of MK.[10,24] A second limitation of this study is that it focuses only on milestone ratings in EM and, thus, results cannot be generalized to other specialties. However, we expect that the analytic approach used here could be applied successfully with milestone data from other specialties. Finally, inferences made from these results may not be generalizable to international medical graduates and DOs because we excluded these groups from this analysis. Given that the pathway to residency and licensure differs from U.S. MD-trained students and residents, we felt international medical graduates and DOs effectively represent a different population and would require a separate analysis.

With these limitations in mind, the present study provides some validity evidence for interpreting USMLE Step 1 and Step 2 CK scores and ACGME EM milestone ratings. The consistency between the rank ordering of the SMEs' judgments and the bivariate empirical correlations provides some evidence that the content represented in Step 1 and Step 2 CK and the EM milestones reflects the constructs that the assessments intend to measure. This type of validity evidence is particularly telling given that the data analyzed come from 2 independent assessment systems based on different assessment conditions. The ACGME milestones are early in their development and use and, as such, any incremental validity evidence based on content enhances their value for meeting the challenge of effectively preparing graduates of graduate medical education programs for safe and effective practice.[25,26] The milestones were designed to be a valid measure of clinical performance, but without data like those reported here, it is difficult to know whether they are meeting this goal. Answering this question relies on an incremental systematic attempt to gather validity evidence, of which this study is a small part.

With respect to the multilevel regression analyses, although the relationships found were positive, the magnitude of these positive effects was small, and the effect of Step 1 score was rendered statistically nonsignificant after controlling for performance on Step 2 CK for all subcompetencies except MK-01. With respect to the USMLE, this suggests that while the content included in Step 1 and Step 2 CK may represent the knowledge, skills, and abilities needed to make decisions about entry into supervised practice, scores may not be practically meaningful with respect to predicting subsequent performance in residency training. This may be especially true for Step 1 given the nonsignificant effects noted above. Overall, this study offers further insight into validity issues for both USMLE scores and ACGME milestone ratings. Still, as is the case with structuring a validity argument for the intended use of any assessment, further study is required to better evaluate and synthesize various sources of validity evidence for both assessment systems.

Acknowledgments: Douglas McGee, MD, Einstein Healthcare Network, Philadelphia, Pennsylvania, president of the Accreditation Council for Graduate Medical Education Emergency Medicine Review Committee, helped to facilitate this study. Members of the National Board of Medical Examiners Emergency Medicine Advanced Clinical Examination Task Force provided predictions for the strength of associations between the United States Medical Licensing Examination Step 1 and Step 2 Clinical Knowledge examination scores and emergency medicine milestone ratings.

Funding/Support: None reported.

Other disclosures: Virginia Commonwealth University receives funding for S.A. Santen outside of this work from an American Medical Association Accelerating Change in Medical Education grant for program evaluation. The other authors do not have relevant financial interests.

Ethical approval: This study was approved by the institutional review board of the American Institutes for Research (AIR EX00490) on August 11, 2019.

S.J. Hamstra was vice president, Milestones Research and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois, at the time of writing, and is now professor, Department of Surgery, University of Toronto, Toronto, Ontario, Canada, and adjunct professor, Department of Medical Education, Feinberg School of Medicine, Northwestern University, Chicago, Illinois; ORCID: https://orcid.org/0000-0002-0680-366X.

M.M. Cuddy is measurement scientist, National Board of Medical Examiners, Philadelphia, Pennsylvania; ORCID: https://orcid.org/0000-0002-5756-9113.

D. Jurich is manager, Psychometrics, National Board of Medical Examiners, Philadelphia, Pennsylvania; ORCID: https://orcid.org/0000-0002-1870-2436.

K. Yamazaki is senior analyst, Milestones Research and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois; ORCID: https://orcid.org/0000-0002-7039-4717.

J. Burkhardt is assistant professor, Emergency Medicine and Learning Health Sciences, University of Michigan, Ann Arbor, Michigan.

E.S. Holmboe is chief and Research, Milestones Development and Evaluation Officer, Accreditation Council for Graduate Medical Education, Chicago, Illinois.

M.A. Barone is vice president, Licensure Programs, National Board of Medical Examiners, Philadelphia, Pennsylvania; ORCID: https://orcid.org/0000-0002-4724-784X.

S.A. Santen is senior associate dean and professor of emergency medicine, Virginia Commonwealth University School of Medicine, Richmond, Virginia; ORCID: https://orcid.org/0000-0002-8327-8002.

References

1 Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(suppl 9):S63–S67.
2 ten Cate O, Carraccio C, Damodaran A, et al. Entrustment decision making: Extending Miller's pyramid. Acad Med. 2021;96:199–204.
3 American Educational Research Association; American Psychological Association; National Council on Measurement in Education. The Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.
4 Kane MT. An argument-based approach to validity. Psychol Bull. 1992;112:527–535.
5 Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49:560–575.
6 Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50:741–749.
7 CampbelL DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959;56:81–105.
8 Beeson MS, Holmboe ES, Korte RC, et al. Initial validity analysis of the emergency medicine milestones. Acad Emerg Med. 2015;22:838–844.
9 Beeson MS, Hamstra SJ, Barton MA, et al. Straight line scoring by clinical competency committees using emergency medicine milestones. J Grad Med Educ. 2017;9:716–720.
10 Hamstra SJ, Yamazaki K, Barton MA, Santen SA, Beeson MS, Holmboe ES. A national study of longitudinal consistency in ACGME milestone ratings by clinical competency committees: Exploring an aspect of validity in the assessment of residents' competence. Acad Med. 2019;94:1522–1531.
11 National Board of Medical Examiners. 2018 Directory: NBME & USMLE Committees & Volunteers. https://www.nbme.org/sites/default/files/2020-01/2018USMLE-NBME-Examination-Committee-Members.pdf. Accessed May 7, 2021.
12 Holmboe ES, Yamazaki K, Nasca TJ, Hamstra SJ. Using longitudinal milestones data and learning analytics to facilitate the professional development of residents: Early lessons from three specialties. Acad Med. 2020;95:97–103.

13 Raudenbush SW, Bryk AS. Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd ed. Thousand Oaks, CA: Sage Publications; 2002.
14 Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York, NY: Cambridge University Press; 2007.
15 United States Medical Licensing Examination. Summary Report and Preliminary Recommendations From the Invitational Conference on USMLE Scoring (InCUS), March 11-12, 2019. https://www.usmle.org/pdfs/incus/incus_summary_report.pdf. Accessed May 7, 2021.
16 Cuddy MM, Dillon GF, Clauser BE, et al. Assessing the validity of the USMLE Step 2 Clinical Knowledge examination through an evaluation of its clinical relevance. Acad Med. 2004;79(suppl 10): S43–S45.
17 Margolis MJ, Clauser BE, Winward M, Dillon GF. Validity evidence for USMLE examination cut scores: Results of a large-scale survey. Acad Med. 2010;85(suppl 10):S93–S97.
18 Norcini JJ, Boulet JR, Opalek A, Dauphinee WD. The relationship between licensing examination performance and the outcomes of care by international medical school graduates. Acad Med. 2014;89: 1157–1162.
19 Cuddy MM, Young A, Gelman A, et al. Exploring the relationships between USMLE performance and disciplinary action in practice: A validity study of score inferences from a licensure examination. Acad Med. 2017;92:1780–1785.
20 Ames SE, Ponce BA, Marsh JL, Hamstra SJ. Orthopaedic surgery residency milestones: Initial formulation and future directions. J Am Acad Orthop Surg. 2020;28:e1–e8.
21 Conforti LN, Yaghmour NA, Hamstra SJ, et al. The effect and use of milestones in the assessment of neurological surgery residents and residency programs. J Surg Educ. 2018;75:147–155.
22 Edgar L, Roberts S, Yaghmour NA, et al. Competency crosswalk: A multispecialty review of the Accreditation Council for Graduate Medical Education milestones across four competency domains. Acad Med. 2018;93:1035–1041.
23 Edgar L, Roberts S, Holmboe E. Milestones 2.0: A step forward. J Grad Med Educ. 2018;10:367–369.
24 Mainous AG III, Fang B, Peterson LE. Competency assessment in family medicine residency: Observations, knowledge-based examinations, and advancement. J Grad Med Educ. 2017;9:730–734.
25 Holmboe ES, Yamazaki K, Edgar L, et al. Reflections on the first 2 years of milestone implementation. J Grad Med Educ. 2015;7:506–511.
26 Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—Rationale and benefits. N Engl J Med. 2012;366:1051–1056.