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

Stanley J. Hamstra, PhD, Kenji Yamazaki, PhD, Melissa A. Barton, MD, Sally A. Santen, MD, Michael S. Beeson, MD, MBA, and Eric S. Holmboe, MD

Abstract

Purpose
To investigate whether clinical competency committees (CCCs) were consistent in applying milestone ratings for first-year residents over time or whether ratings increased or decreased.

Method
Beginning in December 2013, the Accreditation Council for Graduate Medical Education (ACGME) initiated a phased-in requirement for reporting milestones; emergency medicine (EM), diagnostic radiology (DR), and urology (UR) were among the earliest reporting specialties. The authors analyzed CCC milestone ratings of first-year residents from 2013 to 2016 from all ACGME-accredited EM, DR, and UR programs for which they had data. The number of first-year residents in these programs ranged from 2,838 to 2,928 over this time period. The program-level average milestone rating for each subcompetency was regressed onto the time of observation using a random coefficient multilevel regression model.

Results
National average program-level milestone ratings of first-year residents decreased significantly over the observed time period for 32 of the 56 subcompetencies examined. None of the other subcompetencies showed a significant change. National average in-training examination scores for each of the specialties remained essentially unchanged over the time period, suggesting that differences between the cohorts were not likely an explanatory factor.

Conclusions
The findings indicate that CCCs tend to become more stringent or maintain consistency in their ratings of beginning residents over time. One explanation for these results is that CCCs may become increasingly comfortable in assigning lower ratings when appropriate. This finding is consistent with an increase in confidence with the milestone rating process and the quality of feedback it provides.

Milestones, now a core component of accreditation-related data collection, define the expected developmental progression of residents during training.[1] At the time of graduation, the program director must sometimes make difficult decisions in determining residents' competence for those who are on the cusp of entering unsupervised practice. Throughout training, additional judgments on residents' competence must be made, and this can be especially difficult during the first 6 months, when residents are relatively unknown to the program director and teaching faculty. To assist in making the best possible decisions about every resident, all Accreditation Council for Graduate Medical Education (ACGME)-accredited training programs in the United States are required to have a clinical competency committee (CCC) that meets regularly to discuss the progress of each resident and assign milestone achievement ratings relevant to that specialty. Across specialties, residents are typically rated on 22 (range: 12–41) subcompetencies every 6 months.

There are several factors that can potentially affect the validity of milestone ratings, including the feasibility and utility of good assessment tools, the engagement of individual faculty raters in the assessment process, and the deliberation process of the CCC itself when discussing and interpreting the progression of each resident.[2] It is conceivable that CCCs might evolve over time in terms of how they assign milestone ratings, especially in the early years following the adoption of a new system, such as the Next Accreditation System (NAS). Since milestones and the NAS itself are relatively new to the graduate medical education community, it is still unclear to what extent individual CCCs have adopted standardized approaches for generating milestone ratings and whether their processes and criteria for generating those ratings have remained consistent from one year to the next. Since the introduction of the NAS, insight into CCC composition and function has come from a small number of in-depth qualitative studies,[3–9] although none of these have systematically studied factors affecting the variability of milestone ratings over time.

The purpose of this study was to investigate whether CCCs were consistent in applying milestone ratings for beginning learners over time or whether trends of increasing or decreasing milestone ratings exist. To reduce the possible influence of other variables, we chose to focus our analysis on milestone ratings obtained 6 months into the first year of training. To

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

Correspondence should be addressed to Stanley J. Hamstra, 401 N. Michigan Ave., Suite 2000, Chicago, IL 60611; telephone: (312) 755-7037; email: shamstra@acgme.org; Twitter: @stanhamstra.

Acad Med. 2019;94:1522–1531.
First published online June 4, 2019
doi: 10.1097/ACM.0000000000002820

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A693 and http://links.lww.com/ACADMED/A694.

examine consistency, we analyzed data across several years of ratings, using the CCC as our unit of analysis.

Method

Beginning in December 2013, the ACGME initiated a phased-in requirement for reporting milestones among specialties. Data from 2013 to 2016 for all 7 of the phase 1 specialties were available to us for analysis, but there was a great deal of variability between specialties in how the milestones were formulated and reported.[10] Therefore, we chose to focus our initial analysis on data from 3 of the phase 1 specialties—emergency medicine (EM), diagnostic radiology (DR), and urology (UR)—because of the consistency in the milestone reporting protocol between these specialties and to build on early published validity research of milestone ratings.[11–14]

This study was reviewed and approved by the American Institutes for Research. All analyses were conducted using deidentified datasets.

Data

We analyzed all initial (i.e., from the first 6 months of training) CCC milestone ratings for first-year (or year 1) residents from 2013 to 2016 from all EM, DR, and UR programs for which we had data. For EM, we chose to focus our analysis on 3-year, rather than 4-year, EM programs since they constitute the vast majority (128 of 167 programs in 2014–2015) of ACGME-accredited EM programs.[15] We examined data from the same set of 419 programs (123 EM, 178 DR, and 118 UR programs) across the entire 4-year study period, with the number of residents in each cohort as follows: n = 2,838, 2,853, 2,919, and 2,928 for December 2013, 2014, 2015, and 2016, respectively. The code and descriptor for each subcompetency for all 3 specialties are displayed in Appendix 1. For reporting purposes, milestones for each subcompetency are entered on a 10-point rating scale, ranging from 0 to 9. Rating scores of 1, 3, 5, 7, and 9 corresponded to milestone levels 1, 2, 3, 4, and 5, respectively. We considered milestone ratings to be an interval scale for analytic purposes, in accordance with other studies involving large numbers of responses.[16,17] National average in-training or in-service examination results were obtained from each of the specialties' respective boards or specialty societies.

Analytic approach

Although the population of residents entering into these programs changed over the study period, we assumed that the cohorts from each year would not differ substantially from one another in overall competence. For statistical purposes, the CCC—not the individual resident—was the unit of analysis; hence, we examined average milestone ratings within each program across residents per academic year.

To examine any systematic change in milestone ratings over time, we employed a repeated measures statistical model, with a mixture of fixed and random effects. The dependent variable in this model was the average milestone rating across residents within program per academic year for each of the subcompetencies. This model allowed for the detection of signs of increasing leniency, increasing stringency, or consistency within programs across the 4 years. Specifically, we applied a random coefficient multilevel regression model to each subcompetency, with the program-level average milestone rating regressed onto the time of observation (i.e., December 2013, 2014, 2015, and 2016). The intercept and the slope of time of observation were specified as randomly varying coefficients across programs. The slope of time of observation, calculated per program, was the primary variable of interest in this study since this parameter would reflect whether, on average, programs maintained consistency in ratings or became more lenient (i.e., positive coefficient) or stringent (i.e., negative coefficient) over the 4 years. To enhance the interpretability of the model's parameter estimates, we coded time of observation as 0–3 for December 2013–2016 data, respectively. Thus, the intercept in the regression model indicated the national-level average rating by programs in December 2013.

Statistical significance, power, and effect size

We chose a 2-tailed *P* value of .05 as the standard for statistical significance. As this study was an initial investigation into programs' shift in milestone ratings across years, we sought to detect either a significant increase or decrease in average ratings. Similarly, as there were no previous studies from which any expected size of rating shift could be derived, we followed Cohen's convention for determining effect size estimates in education and psychology research.[18] We designed the study to detect a small-to-medium shift in milestone ratings across 4 years; thus, we specified the effect size threshold at $f^2 = 0.07$, in accordance with the random coefficient regression model we used in this study.[19] To detect this size of rating shift with a power of 0.80 or greater, at least 120 programs for each specialty were required for analysis.[18]

Results

The results of the random coefficient regression model are reported in Table 1 and Supplemental Digital Appendixes 1 and 2 (at http://links.lww.com/ACADMED/A693) for each of the 3 specialties included in this study. As mentioned above, the main outcome variable for this study was the slope of milestone ratings over time. The slope values in Table 1 and Supplemental Digital Appendixes 1 and 2 (at http://links.lww.com/ACADMED/A693) represent the average slope of milestone ratings across time of observation, calculated per program; a positive slope indicates increased leniency, while a negative slope indicates increased stringency in assigning milestone ratings over time. Values for intercept and slope are expressed in units of milestone ratings (i.e., milestone levels 1, 2, 3, 4, or 5). Intercept values represent the national average across the included programs at the time of the first milestone rating (i.e., December 2013).

Emergency medicine

The analysis of milestone ratings for EM is based on data from 123 programs over the 4-year period covered by this study, representing 1,378, 1,388, 1,424, and 1,437 residents for December 2013–2016, respectively. The results of the random coefficient regression model are reported in Table 1.

The results for the MK01 subcompetency are not reported in this analysis because the milestone language was modified in 2015, midway through the period covered

Table 1

CCC Milestone Ratings of First-Year Residents (During Their First 6 Months of Training) Over Time for 123 ACGME-Accredited Emergency Medicine Programs, 2013–2016

Subcompetency code	Intercept of milestone ratings at time 0 (SE)	Slope of milestone ratings over time (SE)	Effect size (f^2) for slope	Between-program intercept variance (SE)	Between-program intercept-slope covariance (SE)	Between-program slope variance (SE)	Residual variance (SE)
PC01	1.74 (0.03)[c]	−0.034 (0.013)[a]	0.01	0.11 (0.02)[c]	−0.014 (0.006)[a]	0.010 (0.003)[c]	0.058 (0.005)[c]
PC02	1.81 (0.04)[c]	−0.037 (0.014)[b]	0.01	0.14 (0.02)[c]	−0.021 (0.007)[b]	0.012 (0.003)[c]	0.057 (0.005)[c]
PC03	1.75 (0.04)[c]	−0.040 (0.013)[b]	0.02	0.13 (0.02)[c]	−0.016 (0.006)[a]	0.010 (0.003)[c]	0.059 (0.005)[c]
PC04	1.79 (0.04)[c]	−0.049 (0.014)[c]	0.03	0.12 (0.02)[c]	−0.016 (0.007)[a]	0.009 (0.003)[b]	0.068 (0.006)[c]
PC05	1.63 (0.04)[c]	−0.028 (0.013)[a]	0.01	0.17 (0.03)[c]	−0.021 (0.007)[b]	0.008 (0.003)[b]	0.067 (0.006)[c]
PC06	1.79 (0.04)[c]	−0.054 (0.016)[c]	0.02	0.12 (0.02)[c]	−0.024 (0.008)[b]	0.016 (0.004)[c]	0.075 (0.007)[c]
PC07	1.80 (0.04)[c]	−0.059 (0.014)[c]	0.03	0.15 (0.03)[c]	−0.026 (0.008)[c]	0.011 (0.003)[c]	0.063 (0.006)[c]
PC08	1.73 (0.03)[c]	−0.029 (0.013)[a]	0.01	0.10 (0.02)[c]	−0.012 (0.006)[a]	0.006 (0.003)[a]	0.069 (0.006)[c]
PC09	1.62 (0.03)[c]	−0.019 (0.014)	0.00	0.08 (0.02)[c]	−0.007 (0.006)	0.009 (0.003)[b]	0.077 (0.007)[c]
PC10	1.56 (0.04)[c]	0.001 (0.014)	0.00	0.11 (0.02)[c]	−0.015 (0.007)[a]	0.008 (0.003)[a]	0.083 (0.007)[c]
PC11	1.52 (0.04)[c]	−0.011 (0.014)	0.00	0.10 (0.02)[c]	−0.012 (0.007)	0.008 (0.003)[a]	0.084 (0.008)[c]
PC12	1.69 (0.05)[c]	0.003 (0.018)	0.00	0.18 (0.03)[c]	−0.022 (0.011)[a]	0.016 (0.005)[b]	0.116 (0.010)[c]
PC13	1.58 (0.03)[c]	−0.017 (0.013)	0.00	0.07 (0.02)[c]	−0.003 (0.006)	0.005 (0.003)	0.079 (0.007)[c]
PC14	1.50 (0.04)[c]	0.010 (0.015)	0.00	0.08 (0.02)[c]	−0.009 (0.007)	0.010 (0.004)[b]	0.092 (0.008)[c]
SBP01	1.72 (0.04)[c]	−0.045 (0.016)[b]	0.02	0.16 (0.03)[c]	−0.028 (0.009)[b]	0.013 (0.004)[b]	0.089 (0.008)[c]
SBP02	1.67 (0.04)[c]	−0.028 (0.014)[a]	0.01	0.12 (0.02)[c]	−0.016 (0.007)[a]	0.006 (0.003)[a]	0.083 (0.007)[c]
SBP03	1.75 (0.04)[c]	−0.007 (0.018)	0.00	0.15 (0.03)[c]	−0.014 (0.010)	0.019 (0.006)[c]	0.109 (0.010)[c]
PBLI01	1.75 (0.04)[c]	−0.033 (0.018)	0.01	0.14 (0.03)[c]	−0.033 (0.011)[b]	0.016 (0.006)[b]	0.123 (0.011)[c]
PROF01	1.95 (0.04)[c]	−0.045 (0.014)[b]	0.02	0.11 (0.03)[c]	−0.004 (0.008)	0.003 (0.004)	0.110 (0.010)[c]
PROF02	1.77 (0.05)[c]	−0.015 (0.016)	0.00	0.18 (0.03)[c]	−0.033 (0.010)[b]	0.011 (0.004)[b]	0.102 (0.009)[c]
ICS01	1.87 (0.04)[c]	−0.048 (0.016)[b]	0.02	0.17 (0.03)[c]	−0.020 (0.009)[a]	0.015 (0.004)[c]	0.075 (0.007)[c]
ICS02	1.82 (0.04)[c]	−0.037 (0.014)[b]	0.01	0.12 (0.02)[c]	−0.009 (0.006)	0.010 (0.003)[b]	0.068 (0.006)[c]

Abbreviations: CCC indicates clinical competency committee; ACGME, Accreditation Council for Graduate Medical Education; time 0, December 2013; SE, standard error; PC, patient care; MK, medical knowledge (the results for the MK01 subcompetency are not reported here because the milestone language for emergency medicine was modified midway through the study period); SBP, systems-based practice; PBLI, practice-based learning and improvement; PROF, professionalism; ICS, interpersonal and communication skills.

[a] $P < .05$.
[b] $P < .01$.
[c] $P < .001$.

by this study. Of the 22 remaining EM subcompetencies, 13 showed significant downward trends in ratings over time (i.e., PC01–PC08, SBP01, SBP02, PROF01, ICS01, and ICS02), as indicated by negative slopes over time for these subcompetencies. The effect sizes (f^2) for these ranged from 0.01 to 0.03, which are considered small effect sizes according to Cohen's convention.[18] The remaining 9 subcompetencies showed no significant effect of slope (i.e., no increase or decrease in milestone ratings); most of these had lower starting intercepts compared with the other 13 with significant negative slopes, suggesting less room for decline over time. American Board of Emergency Medicine (ABEM) in-training examination scores showed almost no variance from 2013 to 2016 (i.e., ABEM in-training examination national mean scores were 80, 79, 80, and 80, respectively), consistent with our assumption that there would be no substantial difference in overall competence among beginning residents from 2013 to 2016. Figure 1 and Supplemental Digital Appendix 3 (at http://links.lww.com/ACADMED/A694) show program-level trajectories of mean milestone ratings for each of the EM subcompetencies analyzed in this study. These figures illustrate the general downward trend for many of the subcompetencies and the absence of any upward trends, as well as the individual program variability in assigning milestone ratings to year 1 residents.

Diagnostic radiology

The analysis of milestone ratings for DR is based on data from 178 programs over the 4-year period covered by this study, representing 1,167, 1,168, 1,202, and 1,185 residents for December 2013–2016, respectively. The results of the random coefficient regression model are reported in Supplemental Digital Appendix 1 (at http://links.lww.com/ACADMED/A693).

Eight of the 12 DR subcompetencies showed significant downward trends in ratings over time (i.e., PC01, PC02, MK01, MK02, SBP01, PROF01, ICS01, and ICS02), as indicated by negative slopes over time for these subcompetencies. The effect sizes (f^2) for these ranged from 0.01 to 0.03, which are considered small effect sizes

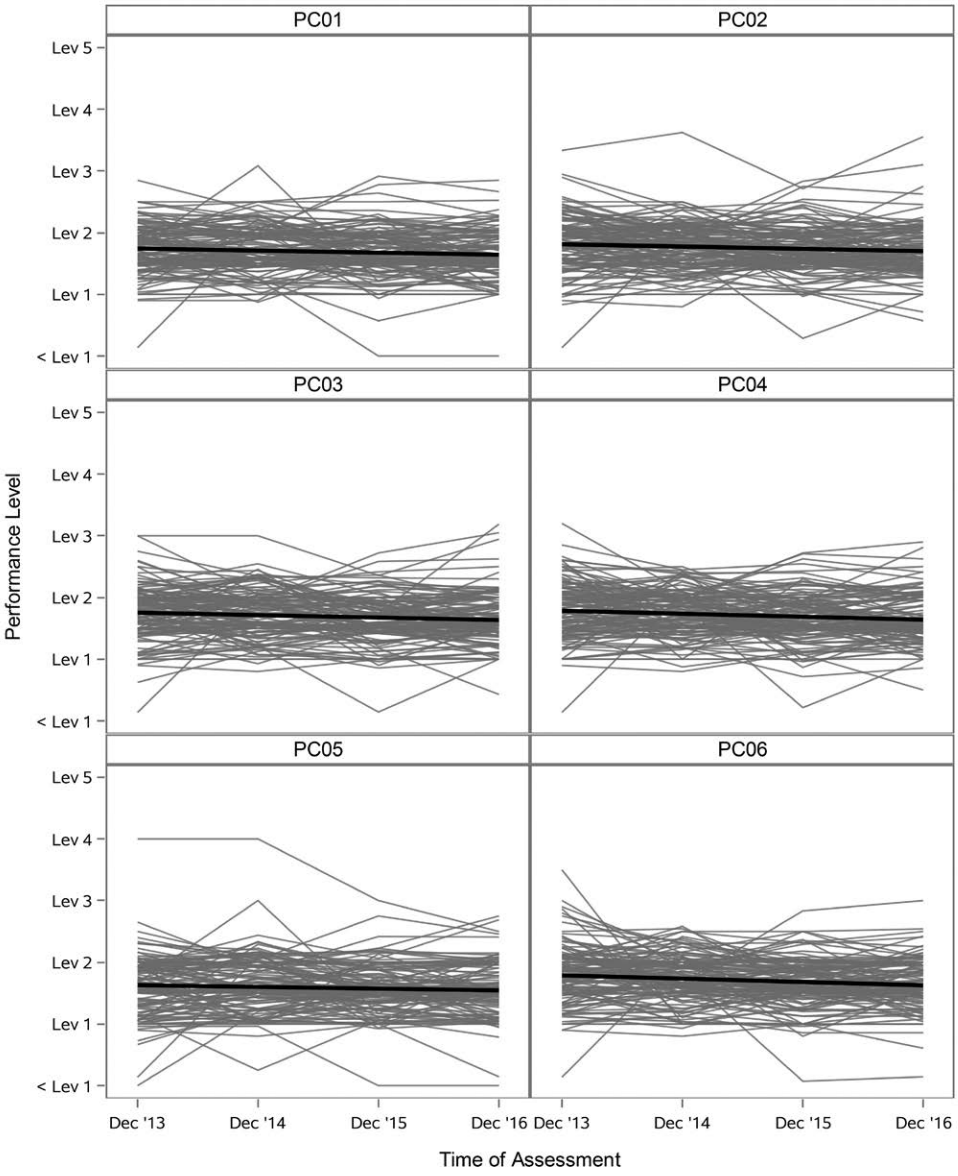

Figure 1 Example program-level mean milestone ratings in EM for year 1 residents, from 123 ACGME-accredited EM programs, 2013–2016 (for the full version of the figure, see Supplemental Digital Appendix 3 at http://links.lww.com/ACADMED/A694). The thick black line is the best-fit regression slope, indicating the national-level trend of program-level average ratings (thin gray lines) over time. The ordinate is scaled in terms of the 5 levels of milestone ratings. The regression line presented here does not account for variations in the number of residents per program. The results for the MK01 subcompetency are not reported here because the milestone language was modified midway through the period covered by this study. Abbreviations: EM indicates emergency medicine; ACGME, Accreditation Council for Graduate Medical Education; PC, patient care; MK, medical knowledge; SBP, systems-based practice; PBLI, practice-based learning and improvement; PROF, professionalism; ICS, interpersonal and communication skills.

according to Cohen's convention.[18] The remaining 4 subcompetencies showed no significant effect of slope (i.e., no increase or decrease in milestone ratings); most of these had lower starting intercepts compared with the other 8 with significant negative slopes, suggesting less room for decline over time. American College of Radiology (ACR) in-training examination scores showed almost no variance from 2013 to 2016 (i.e., ACR in-training examination national mean scores were 53, 54, 53, and 52, respectively), consistent with our assumption that there would be no substantial difference in overall competence among beginning residents from 2013 to 2016. For purposes of illustration, Figure 2 and Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/A694) illustrate the program-level trajectories of mean milestone ratings for all DR subcompetencies, showing a tendency for negative slopes similar to what was observed in EM.

Urology

The analysis of milestone ratings for UR is based on data from 118 programs over the 4-year period covered by this study, representing 293, 297, 293, and 306 residents for December 2013–2016, respectively. The results of the random coefficient regression model are reported in Supplemental Digital Appendix 2 (at http://links.lww.com/ACADMED/A693).

The results for the patient care and medical knowledge subcompetencies

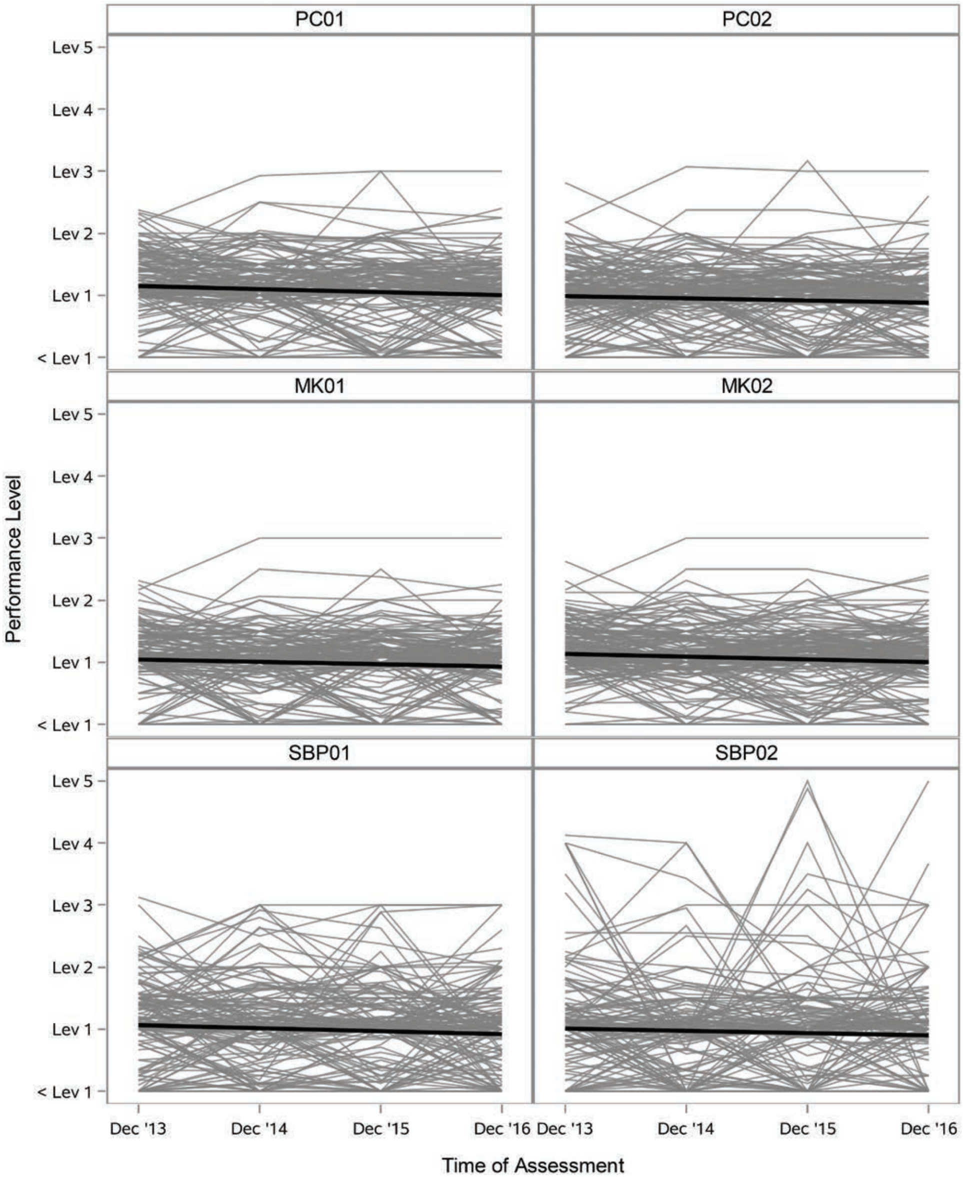

Figure 2 Example program-level mean milestone ratings in DR for year 1 residents, from 178 ACGME-accredited DR programs, 2013–2016 (for the full version of the figure, see Supplemental Digital Appendix 4 at http://links.lww.com/ACADMED/A694). The thick black line is the best-fit regression slope, indicating the national-level trend of program-level average ratings (thin gray lines) over time. The ordinate is scaled in terms of the 5 levels of milestone ratings. The regression line presented here does not account for variations in the number of residents per program. Abbreviations: DR indicates diagnostic radiology; ACGME, Accreditation Council for Graduate Medical Education; PC, patient care; MK, medical knowledge; SBP, systems-based practice; PBLI, practice-based learning and improvement; PROF, professionalism; ICS, interpersonal and communication skills.

are not reported in this analysis because the milestone language was modified in August 2016, before the end of the period covered by this study. Eleven of the 22 remaining UR subcompetencies examined in this study showed significant downward trends in ratings over time (i.e., SBP01, SBP04, PBLI01, PBLI03, PBLI05, PBLI07, PROF02, PROF03, PROF04, PROF06, and ICS04), as indicated by negative slopes over time for these subcompetencies. The effect sizes (f^2) for these ranged from 0.01 to 0.02, which are considered small effect sizes according to Cohen's convention.[18] The remaining 11 subcompetencies showed no significant effect of slope (i.e., no increase or decrease in milestone ratings); most of these had lower starting intercepts compared with the other 11 with significant negative slopes, suggesting less room for decline over time. American Urological Association (AUA) in-service examination scores showed almost no variance from 2013 to 2016 (i.e., AUA in-service examination national mean scores were 55, 56, 56, and 60, respectively), consistent with our assumption that there would be no substantial difference in overall competence among beginning residents from 2013 to 2016. Figure 3 and Supplemental Digital Appendix 5 (at http://links.lww.com/ACADMED/A694) illustrate the program-level trajectories of mean milestone ratings for all UR subcompetencies, showing greater variability than EM or DR but, in general, a similar tendency for negative slopes.

Discussion

Why a gradual shift to stringency matters

This national-level study of milestones data was important in examining the validity of milestone ratings, as well as

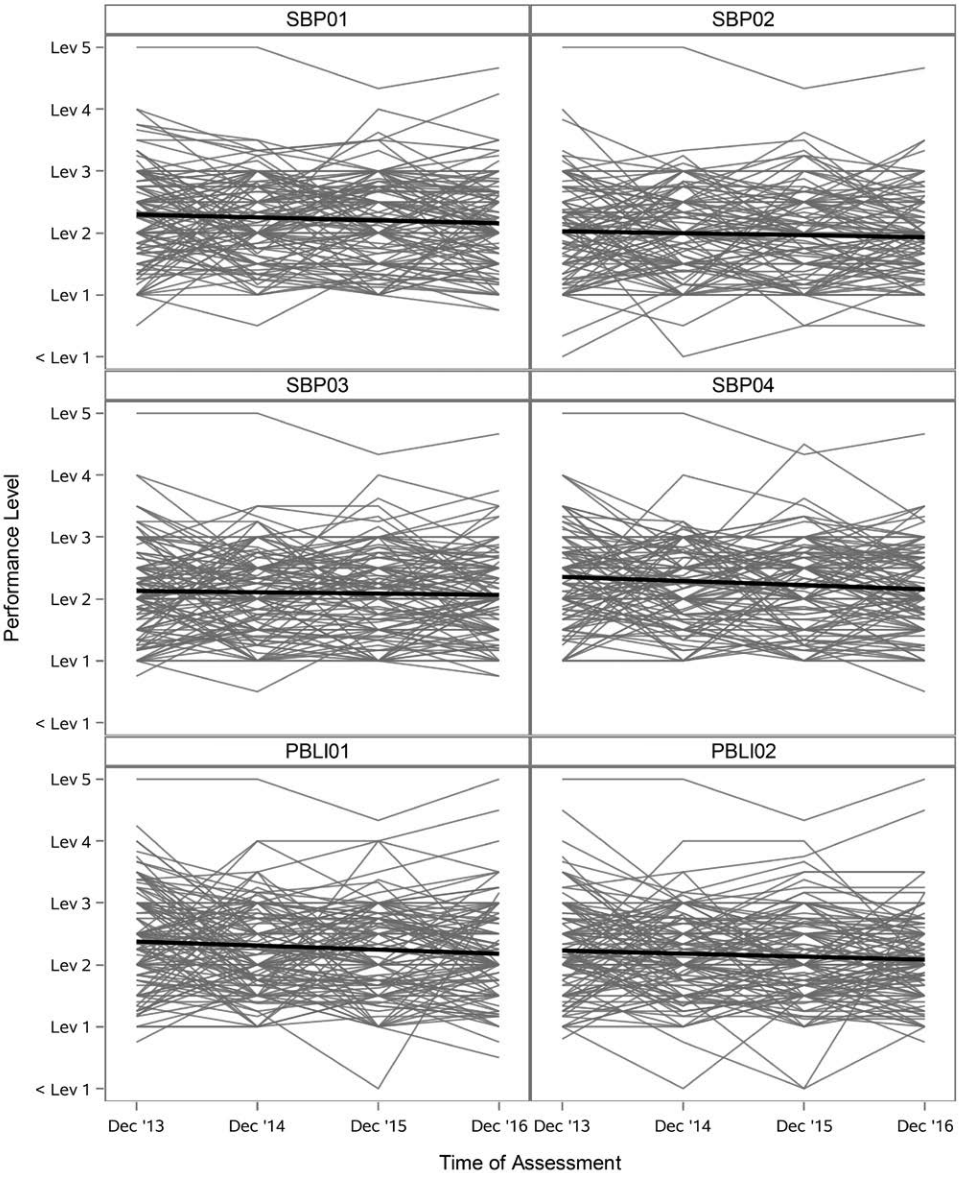

Figure 3 Example program-level mean milestone ratings in UR for year 1 residents, from 118 ACGME-accredited UR programs, 2013–2016 (for the full version of the figure, see Supplemental Digital Appendix 5 at http://links.lww.com/ACADMED/A694). The thick black line is the best-fit regression slope, indicating the national-level trend of program-level average ratings (thin gray lines) over time. The ordinate is scaled in terms of the 5 levels of milestone ratings. The regression line presented here does not account for variations in the number of residents per program. The results for the PC and MK subcompetencies are not reported here because the milestone language was modified before the end of the period covered by this study. Abbreviations: UR indicates urology; ACGME, Accreditation Council for Graduate Medical Education; PC, patient care; MK, medical knowledge; SBP, systems-based practice; PBLI, practice-based learning and improvement; PROF, professionalism; ICS, interpersonal and communication skills.

the possible consequences of certain tendencies in ratings over time. There is substantial preexisting evidence for a tendency to rate learners highly in low-stakes situations (e.g., as with learners at the beginning of their residency).[20] Such a tendency toward leniency might decrease the potential for specific, useful feedback early in training. Medical education in general has a long history of systematic rater errors, such as the halo effect, leniency, and grade inflation.[21,22] The dataset we used offered a unique opportunity to study this phenomenon systematically over time. In contrast to the tendency mentioned above, we found that programs became modestly more stringent in rating first-year residents in more than half of the subcompetencies (32 out of 56) and maintained consistent ratings in the remainder over the 4-year study period. An increase in CCC stringency (i.e., using the lower levels of the milestone scale) in rating first-year residents is consistent with the interpretation of increasing comfort in rating struggling residents lower early in training.[23] The increase in stringency could reflect CCCs' tendency to become more careful (and perhaps vigilant) or more willing to provide room for growth in their ratings of first-year residents as the CCCs gain more familiarity with the milestones framework, its intent, and the quality of feedback it provides.

Increasing comfort with milestones?

Consistency of CCC processes and standards in assigning milestone ratings is especially important for residents early in training because this point in time represents the point of maximum potential for responding to formative feedback. One of the putative benefits of milestones is that they provide a profile of competency achievement that can change over time, so that the learner and the program can adjust educational experiences to maximize learning. We might expect to see less confidence in assigning milestone ratings for residents earlier, rather than later, in training since new residents are typically subject to a high degree of supervision and ratings are made prior to detailed knowledge of the resident's true competence. Such a lack of confidence in the rating scale can lead to grade inflation, especially when faced with possible challenges from residents over low ratings. Milestones are thought to provide an opportunity to counter this tendency by encouraging the use of narrative (criterion-based) descriptors to rate competence rather than numerical ratings based on expected norms for a certain stage in training. Undue emphasis on the numerical aspects of a rating scale, lack of perceived relevance of the narrative descriptors, or ambiguity of narrative descriptors may lead to a lack of certainty in assigning ratings, especially for beginning trainees.[24–26] Over time, this uncertainty could lead to a tendency for a CCC to gradually increase ratings for beginning residents. The results presented here would appear to offer evidence contrary to this expectation and as such represent an encouraging sign that CCCs are making appropriate adjustments to their rating processes. In terms of validity theory, this is consistent with previous evidence of an increase in response process validity with continued used of a psychometric instrument.[27]

Remaining questions

Another potential explanation for these findings is a reduction in uncertainty following the implementation of the NAS in 2013, which might have led to a decrease in variance between programs over time, resulting in the negative slope we observed. Certainly, the between-program intercept–slope covariances in Table 1 and Supplemental Digital Appendixes 1 and 2 (at http://links.lww.com/ACADMED/A693) are an indirect indication that the variation in December 2016 was generally smaller than that in December 2013. A detailed exploration of the reasons for this shift (e.g., using qualitative methods) was beyond the scope of the current study but would be a fruitful direction for future research and could provide further evidence regarding the validity of milestones data. We are currently conducting a number of qualitative studies to examine CCC processes in more depth to expand on this study and earlier work.[3–9]

Some programs did not follow the general tendency for negative slopes, as indicated by the statistically significant between-program slope variance components presented in Table 1 and Supplemental Digital Appendixes 1 and 2 (at http://links.lww.com/ACADMED/A693). This finding is important on its own, as it indicates the potential for the development of national-level resources for each specialty to harmonize their approach to assigning milestone ratings. Finally, consistency in ratings over time can be a proxy for how individual CCCs work. For example, if a program shows variance in ratings for incoming residents over time, then this variance may indicate a need for faculty development to cultivate a shared mental model for discussing and interpreting milestones data.

Limitations

We made several assumptions in developing the analytical model employed in this study. We assumed that the national-level average of first-year resident competence remained stable across academic years (i.e., 2013–2016). While it is possible that we simultaneously witnessed a deterioration in the quality of the entry cohorts over this period, the fact that the ABEM and ACR in-training examination scores showed almost no variance over this time period suggests otherwise. In fact, the AUA in-service examination data even showed a slight increase. However, since we had no other way of independently determining whether the cohorts were equivalent on other dimensions of competence besides medical knowledge, it remains a possibility. This study was restricted to only 3 of the phase 1 specialties for which milestones data have been collected since 2013. Despite our finding of consistency in the results for these 3 specialties, it is possible that these results would not generalize to other specialties. For EM, this study was restricted to 3-year EM programs to maintain homogeneity for analytic purposes. It is possible that inclusion of data from the 4-year EM programs might have yielded different results.

Future directions

Given the demonstrated potential of the analytic approach presented here, our goal is to extend this approach to other specialties to provide further detailed guidance for CCCs in generating valid and meaningful data for program improvement. While these results yielded some insight into the outcome of CCC processes, further study using qualitative methods might help to explain the reasons for these results. The analytic methods developed in this study could then be used to (1) provide information for improving rating processes within programs, (2) modify any subcompetencies with problematic milestone descriptor language, and/or (3) investigate the same question in other specialties. The analytical approach used in this study could be applied every year as a monitoring algorithm to examine patterns of CCCs' judgments over time. In the absence of evidence to the contrary, trends toward increasing leniency could be an impetus for faculty development. As such, the analytical approach described here can be used for the improvement of the CCC's rating processes or potential revision of the narrative descriptors for each subcompetency to better align with the expectations of the program faculty members. Additionally, as other specialties accumulate more milestone assessments over time, the analytic approach in this study can be applied to examine the stability of CCCs' consistency of milestone ratings and provide valuable feedback to programs and CCCs nationally.

One of the advantages of using a random coefficient regression model is that it yields an estimate of intercept and slope for each program. The estimated regression line allows for future investigations to more fully understand program-level deviations from this expected value. For example, quantitative methods can be used to explore the predictive power of potentially meaningful variables such as the number of CCC members, the number of residents within a program, and the variety and type of assessment tools. If these attributes are significant predictors of a program's initial ratings for incoming residents or for the program's rating tendencies over time, then the information could be used for faculty development to alert CCC members and faculty to irrelevant variables that would not be expected to be primary determinants of milestone ratings. Another approach would be to conduct follow-up interviews for programs that show volatile rating patterns over time for incoming residents, which might be a signal of variations in decision-making processes during CCC meetings. With large discrepancies compared with other programs, such data might provide useful insights that could help programs understand how to improve the quality of their rating processes and ultimately make more accurate decisions about preparing residents and fellows for unsupervised practice.

Acknowledgments: The authors wish to thank Dr. Earl J. Reisdorff of the American Board of Emergency Medicine for providing comments on an earlier draft. The authors thank the American Board of Emergency Medicine and the American College of Radiology for providing mean scores for their national in-training examinations and the American Urological Association for providing mean scores for their national in-service examination.

Funding/Support: None reported.

Other disclosures: S.J. Hamstra, E.S. Holmboe, and K. Yamazaki are employed by the Accreditation Council for Graduate Medical Education (ACGME). E.S. Holmboe receives royalties from Elsevier for a textbook on assessment. M.A. Barton is employed by the American Board of Emergency Medicine. M.S. Beeson is director of the American Board of Emergency Medicine. S.A. Santen receives funding from the American Medical Association to evaluate the Accelerating Change in Medical Education grant. None of the investigators received financial compensation for this study.

Ethical approval: This study was reviewed and approved by the American Institutes for Research. All analyses were conducted using deidentified datasets.

Previous presentations: Preliminary results of this work were presented at the Annual Education Conference of the ACGME, Orlando, Florida, March 2017; the annual meeting of the Association for Medical Education in Europe, Helsinki, Finland, August 2017; and the International Conference on Residency Education, Quebec City, Quebec, Canada, October 2017.

S.J. Hamstra is vice president, Milestones Research and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois, adjunct professor, Faculty of Education, University of Ottawa, Ottawa, 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.

K. Yamazaki is senior analyst, Milestones Research and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois.

M.A. Barton is director of medical affairs, American Board of Emergency Medicine, East Lansing, Michigan.

S.A. Santen is professor and senior associate dean, Virginia Commonwealth University School of Medicine, Richmond, Virginia.

M.S. Beeson is director, American Board of Emergency Medicine, East Lansing, Michigan, professor, Department of Emergency Medicine, Northeast Ohio Medical University, Rootstown, Ohio, and program director, Department of Emergency Medicine, Summa Health, Akron, Ohio.

E.S. Holmboe is senior vice president, Milestone Development and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois.

References

1 Nasca TJ, Philibert I, Brigham T, Flynn TC. The Next Accreditation System: Rationale and benefits. N Engl J Med. 2012;366:1051–1056.
2 Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.
3 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.
4 Hauer KE, Chesluk B, Iobst W, et al. Reviewing residents' competence: A qualitative study of the role of clinical competency committees in performance assessment. Acad Med. 2015;90:1084–1092.
5 Ekpenyong A, Baker E, Harris I, et al. How do clinical competency committees use different sources of data to assess residents' performance on the internal medicine milestones? A mixed methods pilot study. Med Teach. 2017;39:1074–1083.
6 Doty CI, Roppolo LP, Asher S, et al. How do emergency medicine residency programs structure their clinical competency committees? A survey. Acad Emerg Med. 2015;22:1351–1354.
7 Schumacher DJ, Michelson C, Poynter S, et al; APPD LEARN CCC Study Group. Thresholds and interpretations: How clinical competency committees identify pediatric residents with performance concerns. Med Teach. 2018;40:70–79.
8 Schumacher DJ, King B, Barnes MM, et al; Members of the APPD LEARN CCC Study Group. Influence of clinical competency committee review process on summative resident assessment decisions. J Grad Med Educ. 2018;10:429–437.
9 Watson RS, Borgert AJ, O'Heron CT, et al. A multicenter prospective comparison of the Accreditation Council for Graduate Medical Education milestones: Clinical competency committee vs. resident self-assessment. J Surg Educ. 2017;74:e8–e14.
10 Sullivan G, Simpson D, Cooney T, Beresin E. A milestone in the milestones movement: The JGME milestones supplement. J Grad Med Educ. 2013;5(suppl 1):1–4.
11 Korte RC, Beeson MS, Russ CM, Carter WA, Reisdorff EJ; Emergency Medicine Milestones Working Group. The emergency medicine milestones: A validation study. Acad Emerg Med. 2013;20:730–735.
12 Beeson MS, Warrington S, Bradford-Saffles A, Hart D. Entrustable professional activities: Making sense of the emergency medicine milestones. J Emerg Med. 2014;47:441–452.
13 Beeson MS, Holmboe ES, Korte RC, et al. Initial validity analysis of the emergency medicine milestones. Acad Emerg Med. 2015;22:838–844.
14 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.
15 Accreditation Council for Graduate Medical Education. Data resource book: Academic year 2014–2015. https://www.acgme.org/About-Us/Publications-and-Resources/Graduate-Medical-Education-Data-Resource-Book. Accessed April 15, 2019.
16 Carifio J, Perla R. Resolving the 50-year debate around using and misusing Likert scales. Med Educ. 2008;42:1150–1152.
17 Leung SO, Wu H. Can Likert scales be treated as interval scales? A simulation study. J Soc Serv Res. 2017;43:527–532.
18 Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
19 Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. A practical guide to calculating Cohen's f(2), a measure of local effect size, from PROC MIXED. Front Psychol. 2012;3:111.
20 Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: Assessor cognition from three research perspectives. Med Educ. 2014;48:1055–1068.
21 Pangaro LN, Durning S, Holmboe ES. Rating scales. In: Holmboe ES, Durning SJ, Hawkins RE, eds. Practical Guide to the Evaluation of Clinical Competence. 2nd ed. Philadelphia, PA: Elsevier; 2018.
22 Kuo LE, Hoffman RL, Morris JB, et al. A milestone-based evaluation system—The cure for grade inflation? J Surg Educ. 2015;72:e218–e225.
23 Holmboe ES, Call S, Ficalora RD. Milestones and competency-based medical education in internal medicine. JAMA Intern Med. 2016;176:1601–1602.
24 Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: Rethinking the etiology of rater errors. Acad Med. 2011;86(10 suppl):S1–S7.
25 Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: Faculty interpretations of narrative evaluation comments. Med Educ. 2015;49:296–306.
26 Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: A quantitative reliability analysis of qualitative data. Acad Med. 2017;92:1617–1621.
27 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006;119:166.e7–166.16.

Appendix 1

Codes and Descriptors for the ACGME Subcompetencies for the 3 Specialties Included in Study of CCC Milestone Ratings of First-Year Residents (During Their First 6 Months of Training) Over Time, 2013–2016

Code	Descriptor
Emergency medicine	
PC01	Emergency stabilization
PC02	Performance of focused history and physical exam
PC03	Diagnostic studies
PC04	Diagnosis
PC05	Pharmacotherapy
PC06	Observation and reassessment
PC07	Disposition
PC08	Task-switching
PC09	General approach to procedures
PC10	Airway management
PC11	Anesthesia and acute pain management
PC12	Goal-directed focused ultrasound
PC13	Wound management
PC14	Vascular access
MK01[a]	Demonstrates appropriate medical knowledge
SBP01	Patient safety
SBP02	Systems-based management
SBP03	Technology
PBLI01	Practice-based performance improvement
PROF01	Professional values
PROF02	Accountability
ICS01	Patient-centered communication
ICS02	Team management
Diagnostic radiology	
PC01	Consultant
PC02	Competence in procedures
MK01	Protocol selection and optimization of images
MK02	Interpretation of examinations
SBP01	Quality improvement
SBP02	Health care economics
PBLI01	Patient safety: Contrast agents, radiation safety, MR safety, sedation
PBLI02	Self-directed learning
PBLI03	Scholarly activity
PROF01	Professional values and ethics
ICS01	Effective communication with patients, families, and caregivers
ICS02	Effective communication with members of the health care team
Urology	
PC01[b]	Gathers information
PC02[b]	Uses diagnostic tests and procedures
PC03[b]	Generates a differential diagnosis
PC04[b]	Develops a patient care plan
PC05[b]	Performs intraoperative and postoperative management of patients
PC06[b]	Performs open surgical procedures
PC07[b]	Performs endoscopic procedures of the upper and lower urinary tracts
PC08[b]	Performs laparoscopic or robot-assisted surgical procedures
PC09[b]	Performs office-based procedures

(*Appendix continues*)

Appendix 1

(Continued)

Code	Descriptor
MK01[b]	Demonstrates level-appropriate competency as indicated by performance on the ABS ITE and AUA ISE
SBP01	Works effectively within and across health delivery systems
SBP02	Incorporates cost awareness and risk–benefit analysis into patient care
SBP03	Works in interprofessional teams to enhance patient safety
SBP04	Uses technology to accomplish safe health care delivery
PBLI01	Improves via feedback and self-assessment
PBLI02	Learns and improves by asking and answering clinical questions from a patient scenario
PBLI03	Acquires the best evidence
PBLI04	Appraises the evidence for validity, impact, and applicability
PBLI05	Applies the evidence to decision making for individual patients
PBLI06	Improves the quality of care for a panel of patients
PBLI07	Participates in the education of other team members
PROF01	Demonstrates adherence to ethical principles
PROF02	Demonstrates compassion, integrity, and respect for others
PROF03	Demonstrates responsiveness to patient needs that supersede self-interest
PROF04	Demonstrates respect for patient privacy and autonomy
PROF05	Demonstrates accountability to patients, society, and the profession
PROF06	Demonstrates sensitivity and responsiveness to diverse populations
ICS01	Communicates effectively with patients and families with diverse socioeconomic and cultural backgrounds
ICS02	Effectively counsels, educates, and obtains informed consent
ICS03	Communicates effectively with physicians, other health professionals, and health-related agencies
ICS04	Communicates effectively during care transitions and consultations with fellow residents
ICS05	Works effectively as a member or leader of a health care team or other professional group

Abbreviations: ACGME indicates Accreditation Council for Graduate Medical Education; CCC, clinical competency committee; PC, patient care; MK, medical knowledge; SBP, systems-based practice; PBLI, practice-based learning and improvement; PROF, professionalism; ICS, interpersonal and communication skills; MR, magnetic resonance; ABS ITE, American Board of Surgery in-training examination; AUA ISE, American Urological Association in-service examination.

[a] The results for the MK01 subcompetency for emergency medicine are not reported in this study because the milestone language was modified midway through the study period.

[b] The results for the PC and MK subcompetencies for urology are not reported in this study because the milestone language was modified before the end of the study period.