Regular Article

Academic Pathology: Volume 8
DOI: 10.1177/23742895211060526
journals.sagepub.com/home/apc

SAGE

Faculty Versus Resident Self-Assessment Using Pathology Milestones: How Aligned Are We?

Sienna Athy, BA[1], Geoffrey Talmon, MD, MEd[2], Kaeli Samson, MA, MPH[3], Kimberly Martin, CTAGME[4], and Kari Nelson, PhD, MA[5]

Abstract

Competent physicians must be able to self-assess skill level; however, previous studies suggest that medical trainees may not accurately self-assess. We utilized Pathology Milestones (PM) data to determine whether there were discrepancies in self- versus Clinical Competency Committee (CCC) ratings by sex, program year (PGY), time of evaluation, and question category (Patient Care, Medical Knowledge, Systems-Based Practice [SBP], Practice-Based Learning and Improvement [PBL], Professionalism [PRO], and Interpersonal and Communication Skills) and Residency In-Service Examination (RISE) score. We completed retrospective analyses of PM evaluation scores from 2016 to 2019 ($n = 23$ residents) 2 times per year. Discrepancies in evaluation scores were calculated by subtracting CCC scores from resident self-evaluation scores. There was no significant difference in discrepancy scores between male versus female residents ($P = .94$). Discrepancy scores among all PGYs were significantly different ($P < .0001$), with PGY1 tending to overrate the most, followed by PGY2. PGY3 and PGY4 underrated themselves on average compared to CCC ratings, with PGY4 having significantly lower self-ratings than CCC compared to any other PGY. In January, residents underscored themselves and in July residents overscored themselves compared to CCC ($P < .0001$ for both). Question types resulted in variable discrepancy scores, with SBP significantly lower than and PRO significantly higher than all other categories ($P < .05$ for both). Increases in RISE score correlated to increases in self- and CCC-scoring. These discrepancies can help trainees improve self-assessment. Discrepancies indicate potential areas for amelioration, such as curriculum adjustments or Milestone's verbiage.

Keywords

Clinical Competency Committee, faculty, Milestones, pathology, residents

Received June 14, 2021. Received revised August 23, 2021. Accepted for publication October 15, 2021.

Introduction

A common method of evaluating progress of trainees is by self-assessment. When self-assessment scores are compared to official evaluations, educators can better identify how accurately a trainee understands his or her skill level. Frequently, evaluation of one's own skill is impacted by cognitive bias known as the Dunning-Kruger Effect, in which lower skilled individuals tend to overestimate their abilities, while experts tend to underestimate their abilities.[1] Accurate self-assessment and skill development to a level of competency and beyond are important characteristics for successful physicians.

[1] Department of Medicine, University of Nebraska Medical Center, Omaha, NE, USA
[2] Department of Pathology and Microbiology and Associate Dean for Medical Education, University of Nebraska Medical Center, Omaha, NE, USA
[3] Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE, USA
[4] Department of Pathology and Microbiology, Omaha, NE, USA
[5] Department of Surgery and Graduate Medical Education Research and Education Program Manager, University of Nebraska Medical Center, Omaha, NE, USA

Corresponding Author:
Kari Nelson, University of Nebraska Medical Center, 985524 Nebraska Medical Center, Omaha, NE 68198, USA.
Email: kari.nelson@unmc.edu

Previous studies regarding self-assessment in medical programs have led to variable results. In one study, it was found that experienced clinicians tended to self-assess more accurately than trainees, but the correlation was not statistically significant. It was concluded from these results that the accumulation of knowledge influences measured competency but may not increase self-assessment skills.[2]

Additional studies have evaluated the validity of resident self-assessment and its correlation with competency. One study examined self-assessment of program/postgraduate year (PGY) 1 residents across 9 different procedural competencies utilizing their objective structured clinical examination (OSCE). They found no significant correlation between the official scoring and resident self-assessment.[3] Another study examined residents in internal medicine to determine levels of skill acquisition through OSCE and self-assessment. Most of the participating residents evaluated themselves lower than their true OSCE score.[4] However, the authors acknowledge that this study was limited by sample size, number of stations, and time allotted for testing.[4] Taken together, the results of these previous studies suggest resident self-assessment may not align with official evaluations.

The Milestone system was developed by the Accreditation Council for Graduate Medical Education (ACGME) to place proper emphasis on skill development in areas important to each respective specialty. In general, all specialty-specific Milestones fall within the 6 ACGME core competencies: Patient Care (PC), Medical Knowledge (MK), Systems-Based Practice (SBP), Practice-Based Learning and Improvement (PBL), Professionalism (PRO), and Interpersonal and Communication Skills (ICS)[5] that allow for longitudinal evaluation and feedback across the continuum of a resident's time in a program.[5] These competencies could be difficult to assess objectively with other traditional testing, so Milestones serve as an added level of evaluation. Although self-assessment using Milestones is not required, it is recommended by the ACGME that pathology residents complete a self-assessment biannually.[6]

One study of the Milestones' efficacy evaluated whether residents could efficiently self-assess their progress using the system.[7] They found that trainees seemed to pinpoint their skill levels more accurately using Milestones than with a general assessment. However, this study did not compare the self-assessment results to faculty evaluations, so the discrepancy between the 2, if any, is unknown.[7]

Starting in 2013, anatomic and clinical pathology residencies utilized the first iteration of 27 specialty specific ACGME Milestones.[5,8,9] A pilot study found that, much like the prior works, the accuracy of self-assessment among trainees was inconsistent; some would consistently underrate while others overrated themselves.[9] Not surprisingly, these investigators determined that utilizing feedback to residents who under- or overrated themselves served to partially correct the discrepancy between the rankings.[9]

The current study aimed to compare residents' self-Milestones scoring to those of the program's Clinical Competency Committee (CCC). This study is novel in that it follows multiple classes of residents over time. We considered these scores relative to the residents' sex, the time of the academic year, PGY, and Resident In-Service Examination (RISE) performance to look for discrepancies and trends. We further examined the specific Milestones categories: PC, MK, SBP, PBL, PRO, and ICS to see whether there were differences between resident and CCC scores within these categories.

Methods

This retrospective study was deemed exempt by the authors' institutional review board. As a part of the semiannual evaluation process, the program's CCC evaluated each resident's progress using the ACGME Milestones (n = 23 residents).[8] The CCC was able to view residents' self-assessments at the time they made their ratings; however, the CCC was blinded from residents' prior self-assessments. Additionally, the CCC was able to view their own previous scores. For the purposes of our study, information on annual RISE performance, PGY, sex, and timing of evaluation (ie, mid-year was January, end of year was July) were collected. All information was deidentified by one researcher with ethical access to the data.

Evaluation of each resident's performance was collected for as many years as the residents worked during the study period (2015-2019). Differences between how residents evaluated themselves versus how their committee evaluated them were calculated by subtracting CCC scores from resident scores (ie, resident − CCC). This value is reported as the "discrepancy evaluation score;" negative scores indicated residents' underestimation of their performance relative to CCC, while positive scores indicated overestimation.

Discrepancies in responses for individual Milestone items were assessed using generalized estimating equations to determine the model adjusted main effects of sex, PGY, month of evaluation, and Milestone category (ie, PC, MK, SBP, PBL, PRO, and ICS). Model results are reported as model adjusted means with associated standard errors and 95% confidence intervals.

To assess associations between evaluations and RISE scores, all raw evaluation data were averaged within each resident, separately for each of their PGYs, and separately for resident and CCC evaluators. These averaged evaluations were used as an outcome variable for general estimating equations, which included the examination score (either raw RISE score or national percentile), sex, PGY, evaluator type (ie, resident or CCC), and an interaction between RISE and evaluator type, to assess whether the association between annual score and evaluation differed by evaluator type. To visualize significant interactions, model estimated means were calculated for each group at the overall minimum and maximum test scores and plotted. *P* values associated with the differences between evaluators at the low and high ends of RISE scores were Bonferroni adjusted. This same analysis between evaluations and RISE scores was repeated using only PC and MK scores for the average calculation of evaluation scores, given these

2 question types most closely align with what the RISE measures. All analyses were performed using SAS software version 9.4 (SAS Institute Inc).

Results

Details of our 23 residents included in the study can be seen in Table 1. Given some residents were ending their program when the study period started, while others were beginning their residency at the end of the study, residents had various lengths of follow-up (Table 1). Specifically, n = 4 residents who were followed continuously for all 4 years.

Based on the adjusted general estimating equation model, there was no significant difference in discrepancy evaluation scores between male versus female residents ($P = .94$, Table 2). However, the month of evaluation was significantly associated with discrepancy ratings ($P < .0001$, Table 2). In January, or roughly halfway through the academic year, residents on average significantly underscored themselves relative to CCC. In July, or at the end of the academic year, residents significantly overscored themselves.

The discrepancy scores among all PGYs were significantly different from one another ($P < .0001$ for all), and one stable trend emerged: PGY1 residents tended to overrate their performance by the largest margin (0.52 [CI: 0.33 to 0.70]), followed by PGY2 (0.16 [CI: −0.03 to 0.34]). Both PGY3 and PGY4 residents underrated themselves on average compared to CCC ratings (−0.23 [CI: −0.42 to −0.05] and −0.42 [CI: −0.61 to −0.23], respectively), with PGY4 having the lowest self-ratings relative to CCC compared to any other PGY (Table 2). This pattern of residents tending to overrate themselves compared to CCC in their early years more so than in later years is visualized in Figure 1.

Discrepancies in ratings were significantly different between Milestone categories in the adjusted model. Specifically, SBP Milestones had significantly lower (−0.21 [CI: −0.39 to −0.04]) discrepancy scores than all other categories (PBLI and MK $P < .05$; PRO, PC, and ICS $P < .0001$), indicating that residents on average underestimated their abilities compared to the CCC in this specific area. The items in the PRO category had a discrepancy score that was significantly

Table 1. Characteristics of Residents Included in the Study.*

Sex	Frequency	Percent
F	10	43.48
M	13	56.52
Follow-up time, years		
1	7	30.43
2	7	30.43
3	5	21.74
4	4	17.39

Abbreviations: F, female; M, male.
*Sex and number of years residents were included in the study are detailed.

Table 2. Model Estimated Means for Discrepancy in Evaluation Scores.*

Model type: Generalized estimating equation						
Outcome: Discrepancy in evaluation scores (resident − committee)						
Variable		Model adjusted means	Standard error	95% Confidence interval		*P* Value
Sex						.94
	F	0.01	0.13	−0.26	0.28	
	M	0.00	0.12	−0.24	0.24	
Program year (PGY)						<.0001†
	1	0.52	0.09	0.33	0.70	
	2	0.16	0.09	−0.03	0.34	
	3	−0.23	0.09	−0.42	−0.05	
	4	−0.42	0.09	−0.61	−0.23	
Month of evaluation						<.0001
	January	−0.07	0.09	−0.25	0.11	
	July	0.08	0.09	−0.11	0.26	
Question type						<.0001‡
	ICS	0.07	0.10	−0.13	0.26	
	MK	−0.05	0.09	−0.24	0.14	
	PBLI	−0.04	0.10	−0.24	0.15	
	PC	0.03	0.09	−0.15	0.21	
	PRO	0.24	0.09	0.06	0.42	

(continued)

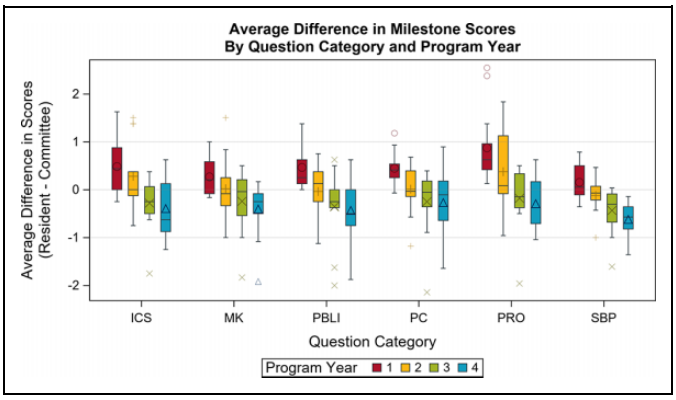

Figure 1. Average discrepancy scores (resident − CCC) per question category on the ACGME Pathology Milestones. A score of 0 indicates that residents and CCC evaluations were the same. There is a notable negative slope for each question category as residents matriculate from program year (PGY) 1 to PGY 4. Symbols indicate outliers. ACGME indicates Accreditation Council for Graduate Medical Education; CCC, Clinical Competency Committee.

higher (0.24 [CI: 0.06-0.42]) than all other categories (ICS $P < .05$; MK, PBLI, PC, and SBP $P < .0001$), suggesting residents have an inflated view of their PRO abilities compared to the CCC rating (Table 2).

When assessing the relationship between national test scores and evaluations, instead of using a discrepancy score, averages of raw evaluation ratings were calculated separately for each resident, PGY, and evaluator type (ie, resident or CCC), and then evaluator type was entered into the model, along with an interaction term between evaluator type and national test score, to be able to assess differences in evaluators. After adjusting for PGY, the association between RISE scores and evaluation ratings significantly differed by evaluator type (CCC vs resident; interaction $P < .0001$; see Figure 2). Specifically, on the lowest end of RISE scores (a score of 397), residents rated themselves significantly higher than CCC ($P = .005$). However, on the highest end of RISE scores (a score of 679), residents rated themselves significantly lower than CCC ($P = .001$).

Discussion

We followed multiple classes of pathology residents over time to determine whether there were discrepancies between resident self- and CCC-ratings using Pathology Milestones. We found that on average residents late in the academic year (ie, July) and early in their program (PGY 1) tended to significantly score themselves higher on Milestones than the CCC did; however, this trend was reversed midway through the academic year (January) and at the end of the program (PGY 4), where residents scored themselves significantly lower than the CCC. Also, residents tended to underestimate their knowledge of SBP relative to all other categories (PBLI, MK, PRO, PC, and ICS) compared to the CCC ratings, while residents tended to overestimate their knowledge of the PRO category relative to all other categories and the CCC. Finally, residents with low

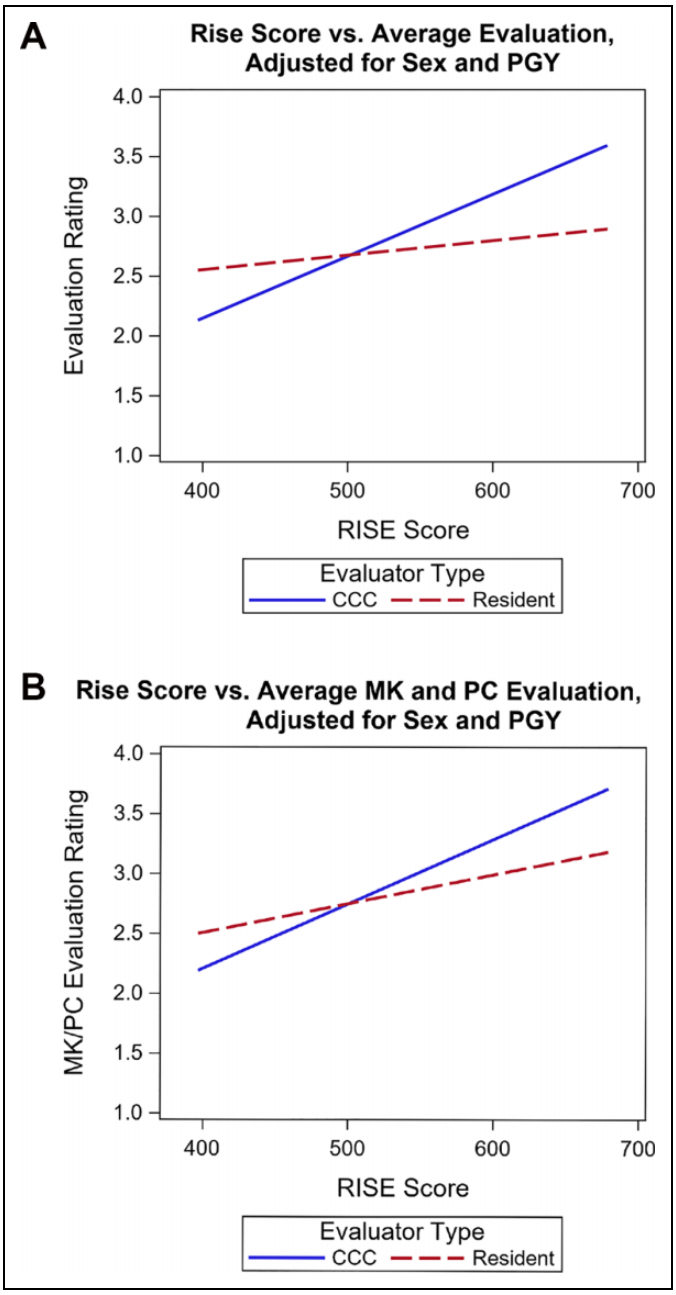

Figure 2. A, After adjusting for PGY and sex, the association between RISE scores and Milestone ratings significantly differed for CCC versus residents (interaction $P < .0001$). Specifically, on the lowest end of RISE scores (a score of 397), residents rated themselves significantly higher than CCC ($P = .005$). However, on the highest end of RISE scores (a score of 679), residents rated themselves significantly lower than CCC ($P = .001$). B, After adjusting for PGY and sex, the association between RISE scores and Milestone ratings, where only MK and PC were used in the Milestone rating average calculation, significantly differed for CCC versus residents (interaction $P = .01$). Specifically, on the lowest end of RISE scores (a score of 397), there was no significant difference in ratings between residents and CCC ($P = .08$). However, on the highest end of RISE scores (a score of 679), residents rated themselves significantly lower than CCC ($P = .03$). CCC indicates Clinical Competency Committee; PGY, program year; RISE, Residency In-Service Examination.

RISE scores tended to significantly overrate themselves compared to CCC, while residents with high RISE scores tended to significantly underrate themselves compared to CCC.

We found that sex did not impact discrepancy score; however, previous studies suggest that sex can impact Milestone ratings. Santen et al found small but significant differences in CCC ratings between males and females, some points that favored males and some favored females.[10] Dayal et al found that males attained higher Milestones ratings as they progressed through residency.[11] Notably, these studies did not compare self-ratings to CCC and based on our findings, there were no differences in discrepancy score between males and females, which suggests that male and female residents may score themselves similar to the CCC and when there is variability, this variability is similar for males and females. However, our finding contradicts a finding in a previous study[12] that examined discrepancy between self- and CCC scoring, which found that female residents were significantly more likely to underscore themselves compared to the CCC than males were. Due to the small sample size of females in the previously published study (n = 7 females)[12] and the current study (n = 10 females), more work needs to be done with a larger sample size to better elucidate this trend.

We also investigated if time of the academic year (January = mid academic year, versus July = end of the academic year) impacted the discrepancy score and found on average that residents overrated themselves compared to CCC at the end of the year but underrated themselves midyear. Our findings corroborate the findings of Lyle et al[12] among surgery programs, which found that self-evaluations tended to be lower than the CCC toward the beginning of the academic year and greater than the CCC at the end of the academic year.[12] Interestingly both studies suggest that self-evaluations within a single academic year do not follow the Dunning-Kruger Effect.

However, when we examined self-evaluations over the entire 4-year residency program, the less experienced PGY1s consistently overrated themselves compared to CCC, while the PGY4s underrated themselves. This was also evident when looking at the individual trajectories of residents (data not shown), where the vast majority followed the same trend and there were no instances of residents who consistently over- or underevaluated themselves over time. This shift over the course of the training program corresponds to the Dunning-Kruger Effect as residents overestimate their abilities early and underestimate their abilities as they develop expertise and become aware of what they do not know.[1] It should be noted that because only 4 residents were followed continuously over all 4 years, the observation of the Dunning-Kruger Effect is primarily based upon comparison of early year residents to later-year residents.

In a previous study among surgery programs, it was found that, apart from PGY3 residents, the trend was for trainees to evaluate themselves lower than the official ranking by a mean of one-half level.[12] Our findings for PGY3-level trainees were similar, but our other PGY years varied from this previous study. Also as noted in previous works, there is debate if residents are "expert enough" in their specialty to evaluate themselves, particularly as this relates to competence.[9,13,14]

Regarding the specific categories within the Milestones, residents tended to underestimate their knowledge of SBP relative to all other categories compared to the CCC ratings. This may be, as many previous studies have documented because SBP (or Health Systems Science [HSS] in other parlance) may not be well understood by the residents and/or may not be clearly taught or prioritized during medical school or residency training. Previous studies have shown that it is often difficult to find educators who have a strong background in these areas, particularly in how to teach SBP/HSS well.[15,16] This lack of teaching or prioritization would likely be the case in the current study if none of the residents had reached a level 4 on Milestones; however, this was not demonstrated in our study, as some residents, but not all, reached a level 4 or higher in this measure by the end of their residency. It is also possible that the residents and the CCC may be using different criteria to evaluate SBP, which would also lead to a discrepancy in scores.

Interestingly, residents tended to overestimate their knowledge of the PRO category relative to all other categories and the CCC. It is well-documented in the literature that training residents and medical students in the areas of professionalism can be difficult.[17,18] Previous reviews of the medical professionalism literature also suggest that it is difficult to measure professionalism due to "frequent use of abstract idealized definitions, the context specific nature of professionalism, and evaluator reluctance to address relatively minor lapses"[19] regarding the Ginsburg et al review.[20] To overcome this, the CCC's assessment of professionalism is very comprehensive as it utilizes 360° evaluations, including input from other faculty, staff, patients, and members of the resident's peer group.

Finally, the positive slopes seen in Figure 2 suggest that both resident self- and CCC evaluation tend to increase on average as RISE score increases. Milestones scores have been shown, in previous studies in surgery[21] and internal medicine,[22] respectively, to correlate to In-Training Exam scores and American Board of Internal Medicine (ABIM) scores, therefore our findings corroborate this previous research. It is important to note, however, that this does not mean RISE scores and all Milestones core competencies have a direct relationship, rather it could be that one, such as MK, or several competencies have a direct relationship to RISE but there may not be a direct relationship between RISE and all core competencies. Interestingly, we do see the general Dunning-Kruger Effect visible in this measure as well because residents who score lower on the RISE tend to overscore themselves relative to the CCC and residents who have the highest RISE scores tend to underscore themselves relative to the CCC.[1] It could be beneficial to both trainees and CCCs to inform them of this effect.

Although this is a unique study of residency Milestones because it follows residents over multiple years of their training program, a limitation of this study is that it involves only a single program at a single institution with a single CCC. Further studies will be needed to determine whether these findings are consistent at other institutions, with other CCCs, or in other

specialties. Additionally, we utilized the CCC ratings as an "expert" group but there may be inherent bias as the CCC members know who the residents are, they are able to see the residents' self-evaluation, and view their RISE score, so this likely impacted their rankings. Additionally, because the CCC was able to view their own previous Milestones scores for the residents as well as the residents' current self-assessment, it should be noted that this may cause confirmation bias.

Another point to consider is that, early in training, the CCC could have difficulty accurately evaluating the trainee due to limited exposure. In fact, early on, the resident may be able to self-assess more accurately than the CCC, so this could be a confounding factor in our study. Although this is unlikely, it should be noted as possible. It is equally possible that residents may push themselves to align their self-evaluation with the CCC, which may be another contributor to the trends we observed. This is a limitation of our study.

Overall, this type of wholistic scoring is typically viewed as more comprehensive, but these limitations may have artificially increased/decreased the CCC's Milestones score or the discrepancy score to some degree and cannot be considered completely independent of one another.

The Milestones are a beneficial part of a wholistic assessment of resident progress over time. Our findings suggest that resident self-evaluation, over time (years) and as content knowledge grows (as measured by RISE score), follows the Dunning-Kruger Effect.[1] As discussed in previous literature, explaining Milestones to the residents at the beginning of their training and utilizing feedback to residents who under- or overrate themselves would be beneficial and may partially correct discrepancies between the rankings.[10]

Conclusions

The ACGME Milestones provide a consistent metric for evaluating trainees' progression through residency. In our program, residents tended to overestimate their abilities compared to CCC early in their training and underestimate their abilities as they completed their final year. Those who scored lower on the RISE also tended to overestimate their abilities. Providing guidance, particularly to early trainees and those who do not perform well on the RISE, may help alleviate some discrepancies.

Authors' Note

Most of these findings were presented at the ACGME annual meeting 2020 in San Diego, CA, as a poster presentation (poster #83). The RISE score calculations/comparisons were added after this presentation in addition to Figure 2B. Sienna Athy contributed to a literature review, interpretation of the data, as well as drafting of/revising the manuscript. Geoffrey Talmon contributed to the conception of the work, interpretation of data, as well as drafting of/revising the manuscript. Kaeli Samson contributed to the analysis of data, drafting of/revising the manuscript. Kimberly Martin contributed to conception of the work and drafting of/revising the manuscript. Kari Nelson contributed to the conception of the work, refinement, and interpretation of data, as well as drafting of/revising the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence led to inflated self-assessments. *J Pers Soc Psychol*. 1999;77:1121-1134. doi:10.1037//0022-3514.77.6.1121
2. Rezaiefar P, Forse K, Burns JK, et al. Does general experience affect self-assessment? *Clin Teach*. 2019;16:197-202. doi:10.1111/tct.12797
3. Barnsley L, Lyon PM, Ralston SJ, et al. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Med Educ*. 2004;38:358-367. doi:10.1046/j.1365-2923.2004.01773.x
4. Tsai JC, Liu KM, Lee KT, et al. Evaluation of the effectiveness of postgraduate general medicine training by objective structured clinical examination—Pilot study and reflection on the experiences of Kaohsiung Medical University Hospital. *Kaohsiung J Med Sci*. 2008;24:627-633. doi:10.1016/S1607-551X(09)70027-0
5. Accreditation Council for Graduate Medical Education, American board of pathology. The pathology milestone project. *J Grad Med Educ*. 2014;6:182-203. doi:10.4300/JGME-06-01s1-09
6. Eno C, Correa R, Stewart NH, et al. *Milestones guidebook for residents and fellows*. Accreditation Council for Graduate Medical Education; 2020;1-19. Accessed November 2, 2019. https://www.acgme.org/Portals/0/PDFs/Milestones/MilestonesGuidebookforResidentsFellows.pdf
7. Bradley KE, Andolsek KM. A pilot study of orthopaedic resident self-assessment using a milestones' survey just prior to milestones implementation. *Int J Med Educ*. 2016;7:11-18. doi:10.5116/ijme.5682.6dfd
8. Naritoku WY, Alexander CB, Bennett BD, et al. The pathology milestones and the next accreditation system. *Arch Pathol Lab Med*. 2014;138:307-315. doi:10.5858/arpa.2013-0260-SA
9. Klutts JS, Guerin LA, Bruch LA, et al. Pathology milestones: assessing clinical competency by committee. *Acad Pathol*. 2015;2: 1-4. doi:10.1177/2374289515614003
10. Santen SA, Yamazaki K, Holmboe ES, Yarris LM, Hamstra SJ. Comparison of male and female resident milestone assessments during emergency medicine residency training: a national study. *Acad Med*. 2020;95:263-268. doi:10.1097/ACM.0000000000002988
11. Dayal A, O'Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training [published correction appears in JAMA Intern Med. 2017 May 1;177(5):747]. *JAMA Intern Med*. 2017;177:651-657. doi:10.1001/jamainternmed.2016.9616

12. Lyle B, Borgert AJ, Kallies KJ, Jarman BT. Do attending surgeons and residents see eye to eye? An evaluation of the accreditation council for graduate medical education milestones in general surgery residency. *J Surg Educ*. 2016;73:e54-e58. doi:10.1016/j.jsurg.2016.07.004
13. Gabbard T, Romanelli F. The accuracy of health professions students' self-assessments compared to objective measures of competence. *Am J Pharm Educ*. 2021;85:8405. doi:10.5688/ajpe8405
14. Valdez CA, Thompson D, Ulrich H, Bi H, Paulsen S. A comparison of pharmacy students' confidence and test performance. *Am J Pharm Educ*. 2006;70:76. doi:10.5688/aj700476
15. Crosson FJ, Leu J, Roemer BM, Ross MN. Gaps in residency training should be addressed to better prepare doctors for a twenty-first-century delivery system. *Health Aff (Millwood)*. 2011;30:2142-2148. doi:10.1377/hlthaff.2011.0184
16. Gonzalo JD, Ogrinc G. Health systems science: the "broccoli" of undergraduate medical education. *Acad Med*. 2019;94:1425-1432. doi:10.1097/ACM.0000000000002815
17. Sullivan C, Murano T, Comes J, Smith JL, Katz ED. Emergency medicine directors' perceptions on professionalism: a Council of emergency medicine residency directors survey. *Acad Emerg Med*. 2011;18:S97-S103. doi:10.1111/j.1553-2712.2011.01186.x
18. Regan L, Hexom B, Nazario S, Chinai SA, Visconti A, Sullivan C. Remediation methods for milestones related to interpersonal and communication skills and professionalism. *J Grad Med Educ*. 2016;8:18-23. doi:10.4300/JGME-D-15-00060.1
19. Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: a review of the literature. *Med Teach*. 2004;26:366-373. doi:10.1080/01421590410001696434
20. Ginsburg S, Regehr G, Hatala R, et al. Context, conflict, and resolution: a new conceptual framework for evaluating professionalism. *Acad Med*. 2000;75:S6-S11. doi:10.1097/00001888-200010001-00003
21. Kimbrough MK, Thrush CR, Barrett E, Bentley FR, Sexton KW. Are surgical milestone assessments predictive of in-training examination scores? *J Surg Educ*. 2018;75:29-32. doi:10.1016/j.jsurg.2017.06.021
22. Hauer KE, Vandergrift J, Hess B, et al. Correlations between ratings on the resident annual evaluation summary and the internal medicine milestones and association with ABIM certification examination scores among US internal medicine residents, 2013-2014. *JAMA*. 2016;316:2253-2262. doi:10.1001/jama.2016.17357