Ability of Ophthalmology Residents to Self-Assess Their Performance Through Established Milestones

Divya Srikumaran, MD,* Jing Tian, MS,† Pradeep Ramulu, MD, PhD,* Michael V. Boland, MD, PhD,* Fasika Woreta, MD, MPH,* Kendrick M. Wang, BS,* and Nicholas Mahoney, MD*

*Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland; and †Wilmer Biostatistics Center, Johns Hopkins School of Public Health, Baltimore, Maryland

OBJECTIVES: Accurate self-assessment is an important aspect of practice-based learning and improvement and a critical skill for resident growth. The Accreditation Council for Graduate Medical Education mandates semi-annual milestones assessments by a clinical competency committee (CCC) for all ophthalmology residents. There are six core competencies: patient care (PC), medical knowledge, systems-based practice, practice-based learning and improvement, professionalism, and interpersonal communication skills. These competencies are assessed by the milestones rubric, which has detailed behavioral anchors and are also used for trainee self-assessments. This study compares resident self-assessed (SA) and faculty CCC milestones scores.

DESIGN: Residents completed milestones self-assessments prior to receiving individual score reports from the CCC. Correlation coefficients were calculated comparing the SA and CCC scores. In addition, statistical models were used to determine predictors of disparities and differences between the SA and CCC scores.

SETTING: Wilmer Eye Institute, Johns Hopkins Hospital.

PARTICIPANTS: Twenty-one residents in the Wilmer Ophthalmology Residency program from July 2014 to June 2016.

RESULTS: Fifty-seven self-assessments were available for the analysis. For each resident's first assessment, SA and CCC scores were strongly correlated ($r \geq 0.6$ and $p < 0.05$) for four milestones, and not correlated for the remaining 20 milestones. In multivariable models, the SA and CCC scores are less disparate for medical knowledge and systems-based practice competencies compared to practice-based learning and improvement. Higher year of training, PC and professionalism competencies were predictive of statistically significant resident overestimation of scores relative to the CCC. In addition, higher CCC scores predicted statistically significant lower SA-CCC disparities and differences. SA-CCC differences did not lower to a significant extent with repeated assessments or modification to the end-of-rotation evaluation forms.

CONCLUSIONS: Self-assessments by ophthalmology residents are not well-correlated with faculty assessments, emphasizing the need for improved and frequent timely feedback. Residents have the greatest difficulty self-assessing their professionalism and PC competency. In general, senior residents and underperforming residents have more inaccurate self-assessments. (J Surg Ed 76:1076–1087. © 2018 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Resident self-assessment, Milestones, Clinical competency committee, Evaluation, Ophthalmology residency, Core competencies

COMPETENCIES: Patient Care, Medical Knowledge, Systems-Based Practice, Practice-Based Learning and Improvement, Professionalism

INTRODUCTION

Given that medicine is a self-regulating profession, there is a critical need for ongoing practice-based learning, self-awareness and development of self-assessment skills. Lack of self-awareness can impair the development of appropriate learning goals and subsequent improvement

Correspondence: Inquiries to Divya Srikumaran, MD, Wilmer Eye Institute, Johns Hopkins University, 1106 Annapolis Road, Suite 290, Baltimore, MD 21113.; e-mail: dsrikum1@jhmi.edu

Abbreviations: ACGME, Accreditation Council for Graduate Medical Education; CCC, clinical competency committee; SA, self-assessed; GEE, generalized estimating equation; PC, patient care; MK, medical knowledge; PROF, professionalism; ICS, interpersonal communications skills; PBLI, practice-based learning and improvement; SBP, systems-based practice.

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2018.12.004

in performance. The Accreditation Council for Graduate Medical Education (ACGME) introduced milestones assessments for residents as a method to improve outcomes-based residency program accreditation.[1] Beginning in December 2014, all ophthalmology residents in ACGME-accredited training programs were required to be evaluated semiannually using a standardized rubric with explicit descriptors of performance. The ophthalmology milestones describe resident performance and skill development for 25 subcompetencies covering the six ACGME core competencies: patient care (PC), medical knowledge (MK), systems-based practice (SBP), practice-based learning and improvement (PBLI), professionalism, and interpersonal communication skills. Though each medical specialty has independently developed milestones rubrics relevant to their appropriate skills set, all accredited residency programs have a clinical competency committee (CCC) composed of at least three faculty members. The CCC reviews all of the existing 360 degree evaluations (patient, faculty, staff and peer) of a resident to grade milestones scores semiannually. One of the proposed benefits of the milestones project is for programs to provide more explicit expectations and feedback to residents, to encourage more informed self-assessment and to aid in early detection of struggling learners.[2] The explicit milestone descriptors also make the tool amenable to resident self-assessment. Performing milestones self-assessments can be a useful method for residents to, gain deeper understanding of expectations, reflect on their performance using a structured format and encourage more informed self-awareness.

Though accurate self-assessment is an important aspect of practice-based learning and improvement, previous studies have suggested that physicians are not skilled at self-assessment and that accurate self-assessment is an unattainable goal.[3,4] Nonetheless as educators we strive to provide feedback to trainees that will inform their self-awareness and learning goals and education. Varying degrees of correlation between resident self-assessed and faculty assessed competency have been reported in other specialties and utilizing different scoring systems.[5–12] However, to our knowledge there is limited literature on self-assessment in ophthalmology[13] and none utilizing milestones assessments. For the milestones assessments, residents and CCC are both examining multisource feedback to determine performance levels.

In this study, we examine the correlation between resident self-assessed (SA) milestones scores with CCC scores and analyze factors that influence disparities and differences between the resident SA and faculty CCC scores in ophthalmology. We hypothesize that there may be a discrepancy between resident self-assessment and CCC assessment and understanding this may help improve resident education. Our primary outcome is how similar or disparate resident SA milestone scores are compared to CCC scores. The secondary outcomes of this study are the factors that contribute to these score disparities and differences. Understanding the factors that impact the discrepancies between faculty and resident self-assessments may in turn shed light on methods to improve resident self-awareness and self-directed learning behaviors as well as influence faculty development in coaching trainees.

MATERIALS AND METHODS

The study protocol was reviewed by the Johns Hopkins University School of Medicine institutional review board and deemed to be exempt. Milestones were assessed as required by the ACGME beginning in December 2014. The full original milestones document is available through the ACGME website (https://www.acgme.org/Portals/0/PDFs/Milestones/OphthalmologyMilestones.pdf). Prior to milestone implementation, the milestones rubric was reviewed and discussed by the Program Evaluation Committee. A subgroup of five faculties from the Program Evaluation Committee then formed CCC scored all residents semiannually using the standard milestones rubric. Faculty members on the CCC met as a group to review, describe, and discuss a plan for assignment of milestone grades. A comprehensive review of all available evaluations from faculty, staff, peers and patients was performed by one member of the CCC, and this faculty member then presented the resident and made recommendations to the committee. Final milestones scores were generated after discussion of the resident's performance and consensus of the committee members. The residents had access to the same multisource feedback evaluations, including written end of rotation and mid rotation faculty reviews, as well as their recollection of informal verbal feedback received on daily assignments and used the same milestones rubric for a self-assessment. The residents each submitted the self-assessment scores to the chair of the CCC prior to receiving their individual score reports which also included a narrative summary of feedback. Each subcompetency has 5 performance levels with written descriptors and the option to score in between levels, thus corresponding to a numerical scale of 0 to 5 in 0.5 increments. The CCC and residents used the same spreadsheet with the milestones transcribed into the file to record all of the milestones scores. The residents did not have access to their current CCC scores and the CCC did not have access to the current resident SA scores when grading the milestones. However, residents who previously had

completed milestones evaluations had their prior performance reports to review. During semiannual review meetings, the program director had both the CCC reports as well as the resident self-assessment available to assist in counseling and highlighting any areas of significant disparity and to help each trainee develop and modify individual improvement goals every 6 months.

The study was performed retrospectively and the residents were not aware that the data would be analyzed for this study while completing the self-assessments. The data for a 2-year period from July 2014 to June 2016 were used for this analysis. The data was collected and compiled by the CCC chair (NM) and the de-identified were given to the statistician (JT) for analysis. Of the 25 ophthalmology milestone ratings that are reported to the ACGME, only 24 were included in the analysis. The MK-1 sub-competency was not consistent across the administered assessments and thus was excluded. The original milestones rubric also includes appendices for different ophthalmic skills that can be separately scored and utilized to help determine at composite score for the PC subcompetencies. These appendix milestones are not separately reported to the ACGME and were not developed by each medical specialty and thus were also not analyzed in this study.

Pearson correlation coefficients (r) were calculated for the SA and CCC scores for the 24 milestones for each resident's first, second, third and fourth assessments, with each assessment considered separately. A Bonferroni correction was performed to adjust for the multiple comparisons. The percentage of residents over- and underestimating their performance was also determined for each milestone at the first assessment. To compare how score differences between SA and CCC vary over time, a mean summary score for each of the six milestone categories was calculated. For example, the summary score for the PC category was calculated by taking the mean of the PC milestones. A paired t test was used to test for significant differences between mean SA and CCC scores for each time point.

Additional models were constructed to determine predictors of SA-CCC difference, calculated by subtracting the CCC score from the SA score. Models employed multiple generalized estimating equations (GEE) to account for within-resident correlations. The analysis was performed using the absolute value of the difference between SA-CCC (|SA-CCC|) to determine predictors of disparities between SA and CCC scores regardless of whether SA was greater than CCC or CCC score was greater than SA. This disparity was analyzed in the GEE model in reference to disparities between SA and CCC of male, first year residents in the PBLI milestone category. The analysis was then also performed using the actual difference between SA-CCC to determine further identify factors predictive of resident overestimation or underestimation of their performance relative to the CCC. Similarly to the GEE model for disparities, a GEE model for differences was created in reference to differences between SA and CCC of male, first year residents in the PBLI milestone category. Specific factors assessed in these models included year of training, milestone category, sex, and the actual CCC scores. An additional model in which year of training was not included, since it is directly linked with the time frame variable, was used to assess the impact of the number of prior milestones assessments and a change in type of end-of-rotation evaluations (midway in the study period the end of rotation evaluations were changed from a 9-point scale without behavioral descriptors to milestones rubric-based descriptive evaluation scale).

RESULTS

Twenty-one residents participated in the study of which 10 were female and 11 were male. Eleven residents had 2 milestones assessments performed while 10 residents had 4 milestones assessments performed during the 2-year study period. Five residents' self-assessments were not completed prior to the residents receiving their CCC milestone score reports and were thus excluded, leaving 57 assessments for analysis. Pearson correlation coefficients with a Bonferroni correction between SA and CCC scores for each of the subcompetencies are shown in Table 1. Sample size was calculated by comparing the paired CCC and SA mean difference. In order to test the significant difference for univariate analysis, including Pearson correlation coefficients (r), at a power level of 0.8, a sample size of $n = 14$ pairs was required. For each resident's first assessment, SA and CCC scores were strongly correlated ($r \geq 0.6$ and $p < 0.05$) for 4 subcompetencies, and not correlated for the remaining twenty. The milestones ratings were highly-correlated for subcompetency in operating room (OR) surgery, MK as applied to PC, use of evidence-based medicine and responsiveness to patient needs. Of these, SA and CCC scores remained strongly correlated for just MK and OR surgery competency scores for 3 of the 4 assessment time points whereas the others were not strongly correlated at any of the subsequent time points as shown in Table 1. When looking at each resident's fourth assessment, the number of competencies with statistically significant strong correlations between SA and CCC ratings did not increase.

Table 2 shows the percentage of first self-assessments by residents that overestimated, underestimated, or matched scores by the CCC. The first self-assessment was used for this analysis since the resident did not have

TABLE 1. Correlation Between Self-Assessment (SA) and Clinical Competency Committee (CCC) Milestones Scores by Assessment Number

Assessment	1 (N = 21)	2 (N = 21)	3 (N = 10)	4 (N = 10)
Patient Care				
PC-1 Patient interview	0.41	0.17	0.24	0.74
PC-2 Patient examination	0.54	0.44	0.30	0.52
PC-3 Office diagnostic procedure	0.24	0.13	***0.68***	0.72
PC-4 Disease diagnosis	0.59	0.34	***0.78***	0.55
PC-5 Nonsurgical therapy	0.49	0.50	0.31	0.54
PC-6 Non- OR surgery	0.47	0.37	0.41	***0.81***
PC-7 OR surgery	***0.64***	***0.63***	0.57	***0.94***
PC-8 Consultation	0.54	***0.65***	***0.85***	0.59
Medical knowledge				
MK-2 Demonstrate level-appropriate knowledge applied to patient management	***0.75***	***0.62***	0.54	***0.92***
Systems-based practice				
SBP-1 Work effectively and coordinate patient care in various health care delivery systems	0.52	0.35	0.10	0.67
SBP-2 Incorporate cost-effectiveness, risk/benefit analysis, and IT to promote safe and effective patient care	0.37	0.58	0.42	−0.09
SBP-3 Work in interprofessional teams to enhance patient safety, identify system errors, and implement solutions	0.08	0.23	−0.04	−0.06
Practice-based learning and improvement				
PBLI-1 Self-directed learning	0.53	0.46	0.63	0.45
PBLI-2 Locate, appraise, and assimilate evidence from scientific studies related to their patients' health problems	***0.62***	0.40	0.35	−0.10
PBLI-3 Participate in a quality improvement project	0.45	0.36	***0.85***	0.58
Professionalism				
Prof-1 Compassion, integrity, and respect for others; sensitivity and responsiveness to diverse patient populations	0.56	0.16	−0.14	0.23
Prof-2 Responsiveness to patient needs that supersedes self-interest	***0.66***	0.39	−0.18	0
Prof-3 Respect for patient privacy and autonomy	0.40	0.41	0.59	0.71
Prof-4 Accountability to patients, society, and the profession	0.30	0.33	0.15	−0.23
Interpersonal communication skills				
ICS-1 Communicate effectively with patients and families with diverse socioeconomic and cultural backgrounds	0.24	0.54	0.10	0
ICS-2 Communicate effectively with physicians, other health professionals, and health-related agencies	0.35	0.40	0.44	***0.87***
ICS-3 Work effectively as a member or leader of a health care team or other professional group	0.33	−0.17	0	−0.10
ICS-4 Effectively present didactic and case-based educational material to physicians and other health care professionals	0.17	−0.08	−0.04	0.51

Statistically significant strong correlations ($p < 0.05$ and $r > 0.6$) are indicated with bold italicized font. A Bonferroni correction was applied given multiple comparisons. PC = patient care, MK = medical knowledge, SBP= systems based practice, PBLI= practice based learning and improvement, Prof= professionalism, ICS= interpersonal communications skills, OR= operating room, IT= information technology.

any prior feedback from the CCC on milestones scores. For each milestone, less than 50% of the residents scored themselves at the same level as the CCC. For the PC milestones, a larger percentage of residents overestimated their skills whereas for MK and SBP milestones a larger percentage of residents tended to underestimate their skills.

To further evaluate these trends over time, a mean summary score for each milestone category was calculated, and the SA and CCC mean summary scores were compared using a paired t test to determine significant differences at each time point. Figure 1 shows the comparison between CCC and SA mean summary scores for each milestone category over time for all residents included in the study at each time point. For PC, the trend lines overlap for the mean summary scores (Fig. 1A) except at the third assessment time point in which the mean SA summary score (3.15) was statistically significantly higher than the mean CCC summary score (2.75) ($p = 0.002$). Figure 1B shows the mean summary score for MK was consistently higher for the CCC relative to SA, and this was statistically significant at the

TABLE 2. Percentage of Residents that Over or Underestimate Milestones Scores Relative to the Clinical Competency in Their First Self-Assessment

	SA = CCC N(%)	SA > CCC N(%)	SA < CCC N(%)
Patient Care Milestones			
PC-1 Patient Interview	5 (23.8)	10 (47.6)	6 (28.6)
PC-2 Patient Examination	2 (9.5)	14 (66.7)	5 (23.8)
PC-3 Office Diagnostic Procedure	6 (28.6)	8 (38.1)	7 (33.3)
PC-4 Disease Diagnosis	4 (19.1)	10 (47.6)	7 (33.3)
PC-5 Non-Surgical Therapy	5 (23.8)	11 (52.4)	5 (23.8)
PC-6 Non- OR Surgery	4 (19.1)	11 (52.4)	6 (28.6)
PC-7 OR Surgery	6 (28.6)	9 (42.8)	6 (28.6)
PC-8 Consultation	9 (42.8)	9 (42.8)	3 (14.3)
Medical Knowledge Milestone			
MK-2 Demonstrate level-appropriate knowledge applied to patient management	6 (28.6)	3 (14.3)	12 (57.1)
Systems Based Practice Milestones			
SBP-1 Work effectively and coordinate patient care in various health care delivery systems	8 (38.1)	6 (28.6)	7 (33.3)
SBP-2 Incorporate cost-effectiveness, risk/benefit analysis, and IT to promote safe and effective patient care	6 (28.6)	4 (19.1)	11 (52.4)
SBP-3 Work in inter-professional teams to enhance patient safety, identify system errors, and implement solutions	6 (28.6)	5 (23.8)	10 (47.6)
Practice Based Learning and Improvement Milestones			
PBLI-1 Self-Directed Learning	7 (33.3)	4 (19.1)	10 (47.6)
PBLI-2 Locate, appraise, and assimilate evidence from scientific studies related to their patients' health problems	8 (38.1)	4 (19.1)	9 (42.8)
PBLI-3 Participate in a quality improvement project	1 (4.8)	13 (61.9)	7 (33.3)
Professionalism Milestones			
Prof-1 Compassion, integrity, and respect for others; sensitivity and responsiveness to diverse patient populations	8 (38.1)	10 (47.6)	3 (14.3)
Prof-2 Responsiveness to patient needs that supersedes self-interest	9 (42.8)	8 (38.1)	4 (19.1)
Prof-3 Respect for patient privacy and autonomy	8 (38.1)	7 (33.3)	6 (28.6)
Prof-4 Accountability to patients, society, and the profession	4 (19.1)	8 (38.1)	9 (42.8)
Interpersonal Communication Skills Milestones			
ICS-1 Communicate effectively with patients and families with diverse socioeconomic and cultural backgrounds	2 (9.5)	8 (38.1)	11 (52.4)
ICS-2 Communicate effectively with physicians, other health professionals, and health-related agencies	6 (28.6)	7 (33.3)	8 (38.1)
ICS-3 Work effectively as a member or leader of a health care team or other professional group	3 (14.3)	9 (42.8)	9 (42.8)
ICS-4 Effectively present didactic and case-based educational material to physicians and other health care professionals	7 (33.3)	6 (28.6)	8 (38.1)

SA = Self -Assessed, CCC =Clinical Competency Committee, PC = Patient care, MK= Medical Knowledge, SBP= Systems Based Practice, PBLI= Practice Based Learning and Improvement, Prof= Professionalism, ICS= Interpersonal Communications Skills, OR= Operating Room, IT= information technology.

fourth assessment time point with mean scores of 3.47 vs 2.73 respectively ($p = 0.006$). The trend lines for both SBP and practice-based learning and improvement overlapped over the study period (Fig. 1C and D). Although the mean summary SA professionalism and interpersonal communications skills milestone scores tended to be higher than CCC, this was not statistically significant (Fig. 1E and F).

To further explore the impact of various factors on the disparity between the SA and CCC scores (|SA-CCC| absolute value of the difference between SA-CCC), a GEE model accounting for sex, year of training, milestone category, and CCC score was utilized (Table 3). For this model, the references used for comparison were the practice-based learning and improvement competency (chosen because of the lowest mean difference between SA-CCC), male sex, and first year of training. Sample size was calculated by analyzing variance in the data. For multivariate models including GEE analysis, a sample size of $n = 62$ pairs was required to include year of training, milestone category and CCC score. To further include gender as a variable in multivariate analysis, a sample size of $n = 399$ pairs was required. Residents in their second year of training had a disparity estimate of 0.1534 which indicates that the average absolute difference between SA and CCC scores for second year

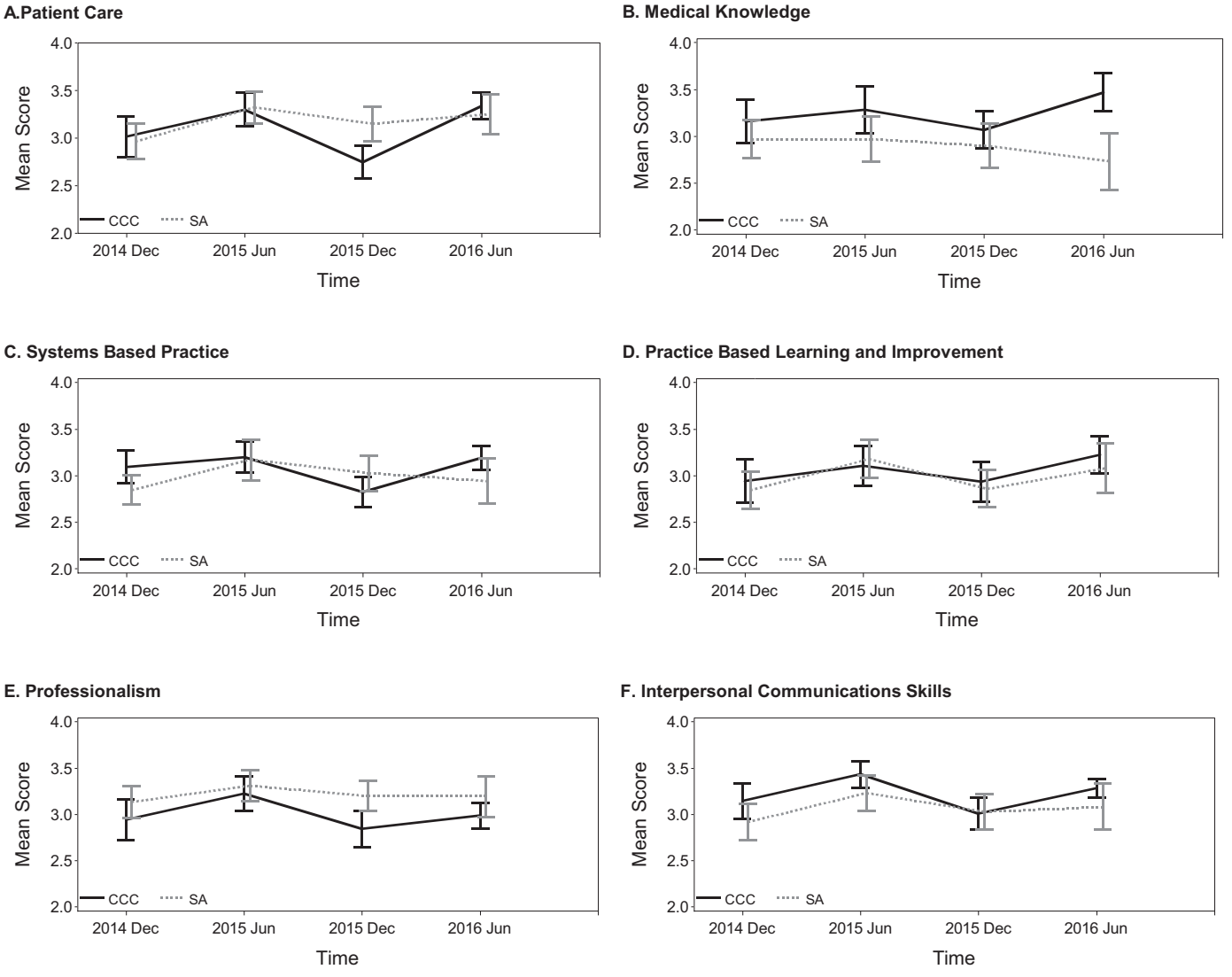

FIGURE 1. Comparison of Mean Clinical Competency Committee (CCC) and Self-Assessment (SA) Summary Scores for the Milestones Categories Over Time. The error bars for each time point represent the standard deviation from the mean summary score.

TABLE 3. Predictors of Disparity (|SA-CCC|) Between Self-Assessment (SA) and Clinical Competency Committee (CCC) Milestones Scores in Relation to a Reference Group

	Estimate	95% Confidence Limits		p Value
Sex (F)	0.1105	−0.0938	0.3148	0.29
2nd Year resident*	0.1534	0.0233	0.2834	0.02
3rd Year resident	0.1946	−0.0458	0.4350	0.11
Milestone category:				
Patient care	−0.0879	−0.2512	0.0753	0.29
Medical knowledge*	−0.2377	−0.4239	−0.0515	0.012
Systems-based practice*	−0.1667	−0.2977	−0.0357	0.013
Professionalism	−0.0901	−0.2328	0.0526	0.22
Interpersonal communications skills	−0.0551	−0.1300	0.0198	0.15
Increased CCC score*	−0.2492	−0.3898	−0.1085	<0.001

Multivariable GEE Model including sex, year of training, milestone category and CCC score to assess predictors of greater disparities |SA-CCC| with the reference group as the practice based learning and improvement milestone category, male sex and first year of training. Estimates indicate a difference in addition to |SA-CCC| for the PBLI milestone of male, first year residents.

* indicates p value <0.05 and is statistically significant.

residents was 0.1534 higher than the absolute difference between SA and CCC scores of the PBLI competency for male first year residents. Thus, residents in their second year of training had greater disparities between SA and CCC scores compared to those in first year of training (p = 0.02). The MK and SBP competencies had less disparities in this model as compared to practice-based learning and improvement competencies suggesting that the residents had more accurate self-assessment of these competencies. In addition, an increase in CCC score, predicted less disparity. In other words, higher performing residents had more accurate self-assessment scores. Sex and the other milestones categories were not significant predictors of disparities between SA and CCC scores. In this model all of the estimates were less than 0.5 which would correspond to the same performance level on the milestones rubric; thus, while findings were statistically significant, the effect size is limited as performance would have been rated at the same level.

To understand factors associated with residents overestimating their skills relative to the CCC, another GEE model containing the same variables was used to determine predictors of the actual difference between SA and CCC (SA-CCC) scores (Table 4). This model again did not show that sex was associated with the degree of difference between SA and CCC scores. However, both second and third year of training were predictive of a higher SA-CCC differences suggesting that senior residents tend to be more likely to overestimate their skills relative to the CCC as compared with first-year residents. The estimates for 2nd year and 3rd year residents are greater than 0.5 and 1.0 respectively corresponding to a half a level of full level of performance higher on the milestones scale. Similarly, the milestone categories of PC and professionalism, were also predictive of greater SA-

TABLE 4. Predictors of Differences (SA-CCC) Between Self-Assessment (SA) and Clinical Competency Committee (CCC) Milestones Scores in Relation to a Reference Group

	Estimate	95% Confidence Limits		p Value
Sex (F)	−0.1650	−0.5937	0.2636	0.45
2nd Year resident*	0.6268	0.4603	0.7933	<0.001
3rd Year resident*	1.0151	0.7889	1.2414	<0.001
Milestone category:				
Patient care*	0.3750	0.2889	0.4611	<0.001
Medical knowledge	−0.1116	−0.3342	0.1109	0.33
Systems-based practice	0.0052	−0.1221	0.1325	.94
Professionalism*	0.2269	0.1061	0.3477	<0.001
Interpersonal communications skills	0.0423	−0.0512	0.1358	.38
Increased CCC score*	−0.8148	−0.8600	−0.7697	<0.001

Multivariable GEE Model including sex, year of training, milestone category and CCC score to assess predictors of greater differences (SA-CCC) with the reference group as the practice based learning and improvement milestone category, male sex and first year of training. Estimates indicate a difference in addition to SA-CCC for the PBLI milestone of male, first year residents.

* indicates p value <0.05 and is statistically significant.

TABLE 5. Change in Differences Between Self-Assessment (SA) and Clinical Competency Committee (CCC) Milestones Scores Over Time

	Estimate	95% Confidence Limits		p Value
Increased number of self-assessments (4 total)	−0.3144	−0.7857	0.1568	0.19
Milestones based end-of-rotation evaluations	−0.5356	−1.2271	0.1560	0.13

Multivariable GEE Model including sex, milestone category, assessment time frame and CCC score to assess predictors of greater (SA-CCC) with the reference group as the practice based learning and improvement milestone category, male sex and the first two milestones assessment.
* indicates p value <0.05 and is statistically significant.

CCC score differences relative to the practice-based learning and improvement category though to a lesser degree since the estimates are less than 0.5. The other milestone categories did not have any statistically significant impact. Finally, as with the model in Table 3, increased CCC scores were predictive of decreased SA-CCC differences ($p < 0.001$).

A separate GEE model containing sex, assessment time frame, and milestone category and CCC score was used to assess the impact of increasing number of assessments and change in our end-of-rotation evaluations (Table 5). Residents who completed a higher number of self-assessments (four total) had lower SA-CCC values compared with residents who only completed two self-assessments; however, the difference was not statistically significant ($p = 0.19$). In addition, midway into the study period, the end-of-rotation resident evaluation form used by faculty was changed from an anchorless Likert scale of 1-9 with 1 corresponding to "Poor" and 9 corresponding to "Exceeds Expectations" to an abridged version of the milestones rubric with similar descriptive behavioral anchors. The SA-CCC difference was also lower after implementation of milestones-based end-of-rotation evaluation forms in our program, but this difference was also not statistically significant ($p = 0.13$).

DISCUSSION

In our program, ophthalmology resident SA and faculty CCC scores were not strongly correlated for a majority of the ACGME subcompetencies assessed. In fact, the SA and CCC scores were only strongly correlated for 4 of the 24 subcompetencies studied for each resident's first assessment. These correlations also did not consistently persist or increase over the subsequent assessments as one might expect even though the residents have more experience with the rubric and the benefit of seeing their prior CCC scores. We also found that senior residents and underperforming residents had greater difficulty with self-assessment and that residents tend to overestimate their PC and professionalism competencies. Our findings are consistent with prior work that suggests that accurate self-assessment is inherently difficult and fraught with tension.[4,14] In our study, our residents and faculty were using the same multisource written evaluation data along with personal experiences and perceptions to inform their ratings. The fact that the residents and CCC often came to different conclusions reinforces the importance of CCC assessments to set expectations, to provide resident feedback and to coach residents to promote greater self-awareness.

Several recent studies have specifically explored the relationship between resident self-assessments and faculty assessments of residents using milestones and other forms of evaluations with varying results in other medical specialties.[5–12] Goldflam et al[5] compared emergency medicine resident self-assessed milestones scores with attending faculty evaluations of residents. They found that the residents consistently rated themselves higher for each subcompetency relative to the mean score of multiple attending faculty. On the other hand, Ross et al[10] found that there was high degree of correlation between resident self-assessed and CCC faculty milestone scores in their anesthesiology residency training program suggesting that the resident self-evaluations might serve as a starting point for CCC discussion. Lyle et al[8] also did not find a high rate of disparity between the resident self-assessments and the clinical competency committee scores in their general surgery residency program. A follow-up multicenter study in general surgery by Watson et al[11] found that there were no major disparities between CCC and resident SA scores in over 70% of the assessments. The differences in the degree of correlation between resident SA and CCC scores for our study and the existing literature may be explained in part by the differing subcompetencies assessed as all specialties developed their milestones rubrics independently. In addition, there may be inherent differences in the programs with respect to evaluations methods and feedback given to trainees. Lyle et al[8] specifically noted that they have a comparable number of faculty and residents with a lot of opportunity for feedback from relatively small number of primary staff. Our program has a very large clinical teaching faculty (greater than 30) and relatively smaller number of trainees ($n = 15$), which might result in greater variability in the quality and quantity of the feedback received by individual residents.

Variations in how the CCC functions within a program and performs their assessments may also impact the degree of correlation between the SA and CCC scores. Currently our CCC members incorporate feedback from all sources of written evaluations for residents including patient, peer, faculty, and support staff. The CCC members also draw on their own interactions with each resident as well as knowledge from other sources of verbal feedback on residents to generate the CCC scores. Differences in the quantity and quality of feedback in each resident's portfolio as well the impact of personal interactions of the CCC faculty with a resident may be a source of bias in CCC scores and a contribution to the reduction in correlation of the SA and CCC scores in our program. Faculty in general have reported more discomfort with giving negative feedback as compared to giving positive feedback directly to the trainees and will often share concerns with the residency leadership rather than directly with the trainee. Though we have conducted faculty development workshops on feedback, we suspect this problem persists. This may have skewed residents' self-perception of their performance and resulted in higher self-assessment scores. Furthermore, the quality and quantity of feedback given to residents on a day to day basis may be different from what faculty report to the CCC through written evaluations in the form of midrotation and end of rotation evaluations. Finally, trainees may be selectively focusing on positive feedback and have greater difficulty accepting and processing negative feedback.

We found that senior resident SA scores had statistically significant differences from the CCC scores and were more likely to overestimate their performance than residents in their first year of training by a half or full level of performance on the milestones rubric. Counter to expectations, although senior residents have received more instances of feedback and previous CCC scores compared to first year residents, the difference between SA scores of senior residents and CCC scores were greater than the difference between SA scores of first year residents and CCC scores. This might be due to increased self-confidence as training progresses, even though actual skills may not have increased substantially. It is also possible that some of our trainees selected the expected milestones level as corresponding to their year of training rather than carefully evaluating the descriptive anchors for each level. One study explored ophthalmology resident self-assessment of cataract surgery skills using a standardized rubric and compared the scores with faculty scores.[13] They found that senior ophthalmology residents were better at self-assessing surgical skills compared to junior residents. Similarly, Goldfam et al[5] also found that senior emergency medicine residents' self-assessments were more similar to the faculty scores compared to junior residents. This is the opposite of our finding and may be related to intrinsic differences in the rubric and skills being assessed and will need further studies to help elucidate the role of year of training.

On average our residents tended to underestimate their MK and overestimate professionalism milestones when we compared average summary scores for these milestone categories over time. However, there was not a statistically significant difference in the mean scores relative to the CCC over the 2-year study period in the univariable analysis. In our multivariable model, the MK and SBP categories actually had lower disparities between SA and CCC scores (|SA-CCC|) compared to the practice-based learning and improvement category, suggesting that residents are more accurate at self-assessing these competencies. In assessing disparities between SA and CCC (|SA-CCC|) in our model, we examined factors that might lead to both over *and* underestimation of SA scores relative to the CCC. We also analyzed the predictors of the differences between (SA-CCC) in a multivariable model to examine when residents tend to overestimate their performance relative to the CCC and found that there is a statistically significant tendency for residents to overestimate their PC and professionalism milestones. However, the effect size of the difference between SA and CCC for the different competencies is less than 0.5 and is not educationally significant since performance levels are reported in 0.5 increments. Lyle et al[8] on the other hand, found that the difference between resident and faculty scores was significant for MK but not the other competencies in their general surgery training program. Further studies are needed to determine if certain domains of competency are more readily self-assessed than others.

We did not find that sex of the trainee influenced the self-assessment. In their subgroup analysis, Lyle et al[8] found that female residents were more likely to accurately self-assess, but when they did have disparity they would also be more likely to underestimate their skills. Watson et al[11] also found that female general surgery residents tended to rate themselves lower than the CCC in their multicenter study. The number of male and female participants in our study is nearly equal, and there was no significant differences or disparities predicted by sex in multiple multivariable models in our study. As stated earlier, the differing results in our study compared to others may be attributable to differences in specialties and other program characteristics.

Another important finding in our study was the impact of the actual CCC score. We found that higher CCC scores predicted lower SA-CCC differences. In other words, residents with lower CCC scores who are underperforming tended to overestimate their skills which might be a sign

of decreased self-awareness and self-reflection. Other studies have also suggested that underperforming residents tend to overestimate their abilities relative to other evaluators and cannot self-identify weaknesses.[6,7] There is even some evidence that residents generate a majority of their learning goals from their own self-assessments, and thus an inability to self-assess accurately may compromise the development of appropriate improvement goals.[15,16] From a program director's perspective, the ability to identify areas in which there are large differences between the SA-CCC scores for individual residents may improve the ability to counsel the resident. Residents with less self-awareness may benefit from additional direct observation and or feedback on a more frequent basis or different methods of feedback delivery that would facilitate acceptance of the feedback and modification in learning plans.[17] Strategies that improve self-awareness by coaching residents to seek and assimilate external feedback may in turn facilitate remediation efforts for underperforming residents. We have tried to encourage residents to take a more active role in seeking feedback especially when there are performance concerns. We have also hosted faculty development workshops in an attempt to increase faculty comfort and skills in providing explicit verbal and written feedback to residents, especially when the feedback is negative. The residents performed multiple self-assessments over time with feedback from the CCC and program director between those assessments. Each resident received a letter summarizing their milestones score report as well as a graph showing the de-identified scores of all of the residents in the program for each milestone. In addition to a required meeting with the program director every 6 months, the residents were also encouraged to meet with the CCC chair if any there was anything unexpected on the CCC report. Although the SA-CCC difference was also lower for residents with a greater number of self-assessments this finding was also not statistically significant. Further research to identify optimal quantity, quality, and method of delivering feedback could be beneficial to decreasing the difference between self-assessed performance and CCC assessed performance. Nonetheless, we feel that the approach of having resident self-assess and compare their ratings to the CCC can be very informative for the trainee and the program director and can be utilized by all specialties.

There are several limitations to this study. First, the single center nature of this study limits generalizability to other residency programs due to variations in program size, faculty-resident ratios, feedback quality and quantity, and evaluation methods. Our residency program has a high number of faculty members relative to a low number of residents, and this may not be representative of other ophthalmology residency programs. Second, the small sample number of this study limited ability to detect statistically significant differences. Based on our sample size calculations, we are sufficiently powered to detect differences for univariate analysis; however, we are not adequately powered to detect a significant difference for multivariate analysis. Third, the specific findings relating to each individual subcompetency may not be able to be generalized across other specialties since subcompetencies differ by specialty. However, the process of comparing resident self-assessed and faculty/CCC ratings can be generalized to all programs since all ACGME accredited residency programs are required to report milestones based on the 6 core competencies, and the 2017 ACGME Milestone Guidebook for Residents and Fellows specifically encourages residents and fellows to perform self-assessments and compare with the CCC ratings to enhance self-directed learning and awareness.[18] Fourth, another limitation is the relatively short follow-up time of 2 years in this study. Having additional years of follow-up on a trainee with 6 milestones assessment might yield different results.

During the study period, we changed our end-of-rotation evaluation forms to mirror the milestones descriptors. This change made the language of the feedback tool used by faculty at the end of each rotation the same as the language they used for the self-assessment. This change resulted in lower SA-CCC difference however not to a statistically significant extent. It is possible that the relatively short study period and relatively small number of residents limited our ability to detect a significant difference after implementations of these changes. Future prospective multicenter studies utilizing a larger sample size and longer follow-up may help elucidate methods to enhance informed resident self-assessment and the disparities between SA and CCC scores. While a larger sample size from a multicenter study may allow smaller differences to be appreciated and increase generalizability among residency programs, the increase in variability arising from including multiple centers would need to be accounted for in sample size calculations. However, assuming that variability will remain similar to that observed in this study, a sample size of $n = 62$ is needed for multivariate analysis examining year of training, milestone category, and CCC score, and a sample size of $n = 399$ is needed for multivariate analysis examining sex, year of training, milestone category, and CCC score. With larger sample sizes, future multicenter studies may also be able to analyze more trainee covariates such as demographics, gender, age, year of training and other program characteristics such as faculty to resident ratio, evaluation, and assessment methods.

CONCLUSION

In conclusion, SA and CCC scores were not highly correlated overall in this study and residents tend to overestimate their PC and professionalism competencies in relation to faculty assessments. In addition, senior residents and underperforming residents have less accurate self-assessments. Our study contributes further to the existing literature on self-assessment by identifying factors at a single institution that impact the disparities and differences between SA and CCC assessments. This study includes a single mid-sized ophthalmology residency program and thus may limit generalizability to other specialties and programs. However, completion of milestones self-assessment provided the residents an opportunity for structured reflection on their performance. The identification of major disparities between SA and CCC assessments can be helpful in counseling individual residents.

DECLARATIONS

Ethics Approval: The study was submitted to and reviewed by the Johns Hopkins Institutional Review Board and deemed to be exempt (IRB00103502). A waiver of consent was granted because the study involved no more than minimal risk to subjects and the waiver would not adversely affect the rights and welfare of the subjects.

Consent for Publication: not applicable

Availability of data and materials: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing Interests: The authors declare that they have no competing interests.

Funding: The statistical analysis was supported by the Wilmer Biostatistics Core Grant EY01765.

Author Contributions: DS, PR, FW, MB and NM are all members of our clinical competency committee and contributed to the study design, acquisition of data, and interpretation of the results and contributed to the writing of the manuscript. JT performed the statistical analysis. KW contributed to writing the manuscript. All authors critically reviewed and approved the final manuscript.

Author's Information: DS, PR, FW, MB, and NM are all members of our clinical competency committee and heavily involved in our residency training program. DS is the vice chair of education and former residency program director during the study period. FW and NM are associate residency program directors. MB is the current residency program director. PR is the current glaucoma division chief.

Meeting Presentation: This work was previously presented in part as a poster at the Educating the Educators Meeting, Association of University Professors of Ophthalmology, San Diego, 2017.

REFERENCES

1. Lee AG, Arnold AC. The next accreditation system in ophthalmology. *Surv Ophthalmol*. 2015;60:82–85.
2. Holmboe ES, Edgar L, Hamstra S. The Milestones Guidebook. ACGME; 2016.
3. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA*. 2006;296:1094–1102.
4. Eva KW, Regehr G. "I'll never play professional football" and other fallacies of self-assessment. *J Contin Educ Health Professions*. 2008;28:14.
5. Goldflam K, Bod J, Della-Giustina D, Tsyrulnik A. Emergency medicine residents consistently rate themselves higher than attending assessments on ACGME milestones. *West J Emerg Med*. 2015;16:931–935.
6. Gow KW. Self-evaluation: how well do surgery residents judge performances on a rotation? *Am J Surg*. 2013;205:557.
7. Lipsett PA, Harris I, Downing S. Resident self-other assessor agreement: influence of assessor, competency, and performance level. *Arch Surg (Chicago, Ill.: 1960)*. 2011;146:901–906.
8. 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:e58.
9. Meier AH, Gruessner A, Cooney RN. Using the ACGME milestones for resident self-evaluation and faculty engagement. *J Surg Educ*. 2016;73:e157.
10. Ross FJ, Metro DG, Beaman ST, Cain JG, Dowdy MM, Apfel A, et al. A first look at the Accreditation Council for Graduate Medical Education anesthesiology milestones: implementation of self-evaluation in a large residency program. *J Clin Anesth*. 2016;32:17–24.
11. Watson RS, Borgert AJ, O Heron CT, Kallies KJ, Sidwell RA, Mellinger JD, 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.

12. Abadel FT, Hattab AS. How does the medical graduates' self-assessment of their clinical competency differ from experts' assessment? *BMC Med Educ*. 2013;13:24.

13. Casswell EJ, Salam T, Sullivan PM, Ezra DG. Ophthalmology trainees' self-assessment of cataract surgery. *Br J Ophthalmol*. 2016;100:766–771.

14. Sargeant J, Mann K, van der Vleuten C, Metsemakers J. "Directed" self-assessment: practice and feedback within a social context. *J Contin Educ Health Professions*. 2008;28:47.

15. Bounds R, Bush C, Aghera A, Rodriguez N, Stansfield RB, Santen SA, et al. Emergency medicine residents' self−assessments play a critical role when receiving feedback. *Acad Emerg Med*. 2013;20:1055–1061.

16. Plant J, Corden M, Mourad M, O'Brien B, van Schaik S. Understanding self-assessment as an informed process: residents' use of external information for self-assessment of performance in simulated resuscitations. *Adv Health Sci Educ*. 2013;18:181–192.

17. French JC, Colbert CY, Pien LC, Dannefer EF, Taylor CA. Targeted feedback in the milestones era: utilization of the ask-tell-ask feedback model to promote reflection and self-assessment. *J Surg Educ*. 2015;72:274.

18. Jardine D, Deslauriers J, Kamran SC, Khan N, Hamstra S, Edgar L. Milestones Guidebook for Residents and Fellows. ACGME; 2017.