Assessing Progression of Resident Proficiency during Ophthalmology Residency Training: Utility of Serial Clinical Skill Evaluations

Grace L. Paley1, Thomas S. Shute1, Geetha K. Davis2, and Susan M. Culican*1
1Department of Ophthalmology and Visual Sciences, Washington University, St Louis, MO, USA
2Department of Ophthalmology, University of Missouri, Columbia, MO, USA
*Corresponding author: Susan M. Culican, Email: culican@wustl.edu
Received: 12 June 2017; Accepted: 01 September 2017; Published: 09 September 2017

Abstract

Objective: The Accreditation Council for Graduate Medical Education (ACGME) has mandated that residency programs document progression of competency-based outcomes. The Ophthalmic Clinical Evaluation Exercise (OCEX) assesses clinical skills in ophthalmology residents during patient encounters. Although OCEX has been validated for assessing several of the ACGME-mandated competencies, it was unclear whether OCEX can measure the development of proficiency during residency. This study evaluated whether OCEX can discriminate skill levels across years in training.

Methods: In 2017, the authors performed a retrospective analysis of modified OCEX evaluations collected for 22 residents over a 3-year residency program. OCEX subcomponent scores were averaged to generate a mean score for each evaluation, followed by linear regression analysis for mean scores over time for individual residents. One-way ANOVA with repeated measures was used to compare regression analysis for mean scores over time for individual residents. OCEX subcomponent scores were averaged on modified OCEX evaluations collected for 22 residents over 3 years at 2 residency programs. OCEX subcomponent scores were averaged to generate a mean score for each evaluation, followed by linear regression analysis for mean scores over time for individual residents. One-way ANOVA with repeated measures was used to compare regression analysis for mean scores over time for individual residents.

Results: Mean OCEX scores for individual residents and resident class averages showed variable trajectories over the course of residency. There was no consistent effect of increasing level of training on scores. Surveys of evaluators and program directors indicated different interpretations of the grading scale anchors and irregular participation by faculty.

Conclusion: This dataset suggests that, despite clear behavioral anchors and faculty development on the use of the tool, evaluators still apply inconsistent grading standards that limit the OCEX from accurately monitoring longitudinal development of resident clinical performance in real-world practice.

Introduction

Across all postgraduate medical training programs, a variety of clinical evaluation exercise (CEX) feedback forms are used to monitor and promote resident education. CEX and the shorter “mini-CEX” provide structured scoring rubrics for assessing a resident’s performance during an observed resident-patient encounter [1,2]. Typically these forms are completed by an attending physician, who grades the resident on how well he or she conducted a history and physical exam and formulated an assessment and plan. When the Accreditation Council for Graduate Medical Education (ACGME) mandated documentation of resident performance, program directors readily utilized CEX forms to help fulfill this requirement [3]. By taking these “snapshots” of residents in action, programs could simultaneously emphasize the acquisition of specific clinical competencies, provide residents with immediate formative feedback, and foster earlier detection of struggling or non-progressing residents, thus allowing for earlier intervention.

For many ophthalmology residency programs in the United States, the cognate evaluation form is usually the Ophthalmic Clinical Evaluation Exercise (OCEX) [4]. The OCEX provides a checklist of competencies under the section headers of interview skills, ophthalmologic examination, interpersonal skills and professionalism, and case presentation. Each competency is individually graded on a rating scale ranging from 1-4, correlating to performance that is below expectations, meets some expectations, meets all expectations, or exceeds expectations, respectively. The expectations are based on the capabilities of a competent resident at the time of graduation, such that serial OCEX evaluations may be employed to show improvement in clinical competencies as a trainee progresses through residency [4,5]. Using a videotaped resident-patient encounter graded by nearly a hundred ophthalmology teaching faculty nationwide, the OCEX was shown to be a reliable and valid tool for assessing resident clinical skills and behavior [4,5]. To the best of our knowledge, however, we know of no evidence that OCEX reliably captures the development of resident skills over time in residency as originally intended. The ability to demonstrate longitudinal improvement in clinical competencies is key to meeting the ACGME mandate.

To address whether OCEX is a useful metric of resident clinical skill progression, we performed a retrospective analysis on OCEX evaluations collected for 22 resident physicians over a 3-year residency program at 2 institutions. We hoped to see an overall positive trend in performance ratings as novice residents tackled the learning curve, with an eventual plateau in the final year as the residents achieved clinical expertise in preparation for independent practice. Specifically, we predicted that scores would improve over time for individual residents relative to their stage of training, and correspondingly, that scores should be higher for residents in their final year compared to those in their first year.

Here we report our experience that, while the aggregate data suggested improvement in OCEX scores over time, individual residents and resident class averages did not uniformly show positive score trajectories over the course of residency. Additionally, analysis of individual evaluator scoring trends and results of a faculty survey both indicated that faculty evaluators interpret OCEX grading anchors differently and thus apply inconsistent OCEX grading scales to the same residents. These findings invite discussion regarding our current best practices for evaluating resident performance in terms of what they do well and how they might be improved to better capture the intended data.

Methods

The modified OCEX form includes 19 competencies ranging from ophthalmic technical skillsets, such as visual acuity or direct ophthalmoscopy, to general categories, such as communication skills or professionalism (Supplemental Figure 1). Competencies are scored on a scale of 1 to 4 based on the meeting of expectations, facility with exam technique, and ability to detect exam findings. Each competency has a short written description (grading anchor) indicating qualities that justify each numeric score. There is also a section for freehand comments. Since a number of evaluations did not include scores for all 19 competencies, a summative mean OCEX score was determined for each evaluation. If one or more competencies were left blank on the form, the mean score was adjusted to reflect the actual number of
In this study, OCEX evaluations were retrospectively analyzed for 22 residents from 2 ophthalmology residency programs over 3 years of training (postgraduate year (PGY)-2 through PGY-4). With one exception, residents received at least one and usually multiple OCEX evaluations per year. Evaluations were completed at the discretion of the faculty, and residents tended to have more evaluations at earlier timepoints in residency. A single faculty member completed the majority of evaluations, although 25 total evaluators were involved across both programs. Resident files were de-identified prior to data extraction. An institutional review board (IRB) exemption was obtained for this medical education study (Washington University IRB# 201702096; University of Missouri IRB# 208053). Linear regression, one-sample t-tests, and one-way ANOVA with repeated measures were used to examine OCEX score trends over time. Statistical significance was defined as p-value <0.05. Analyses were performed in SPSS version 22 (IBM SPSS Statistics, Armonk, NY).

### Results

The mean OCEX score was plotted against time in residency for individual residents with linear regression analysis (Figure 1). Surprisingly, the score trends were quite variable across the residents and only a handful of individuals displayed clinical improvement over time as suggested by an overall positive incline. For some residents, long periods of time elapsed between successive evaluations especially as graduation approached, leading to large jumps in the plots. In order to minimize the impact of changes in the pool of faculty graders over the years, the residents were separated into 5 classes by graduation year and the mean OCEX score trends analyzed by class. Using linear regression of mean OCEX score averaged across each class over the 3-year residency, 2 classes showed a statistically significant improvement (p=0.04, 0.005), 1 class showed a statistically significant decline (p=0.02), and 2 classes showed no significant change (p=0.06, 0.11; one-sample t-test compared to a theoretical mean slope of zero). The erratic visual impression of the individual plots, despite taking a mean score for each clinical evaluation across multiple competencies, suggests that either resident physicians did not improve in clinical skills.

![Mean OCEX score over time for individual residents](image)

**Figure 1:** Mean OCEX score over time for individual residents. Scatter plots of Mean OCEX evaluation scores over time spent in residency for individual trainees. Each data point represents a single OCEX evaluation. The graphs are arranged in order of the slope of the linear regression lines, demonstrating the variable trends across the dataset.

over the course of residency training, or that the evaluation tool did not adequately measure the expected progression of these skills over time.

In order to better account for time gaps during which residents received fewer evaluations, we also looked at mean OCEX scores binned by year of training. We hypothesized that the mean OCEX score should increase across 3 successive years of training as residents developed their clinical skills. Specifically, for each resident, scores for each individual competency on OCEX were averaged across an academic year to generate a mean score for the postgraduate year (PGY-mean). Of the 19 competencies, 3 ophthalmic skillsets (ocular motility, gonioscopy, and direct ophthalmoscopy) were assessed less frequently and not every year. Thus, these 3 competencies were omitted from the analysis binned by year. For each resident, the PGY-means for each of the 16 remaining competency scores were then averaged to generate an overall OCEX score for each postgraduate year. One resident did not have any OCEX evaluations as a PGY3, so this resident was excluded from this analysis. The mean OCEX score aggregated by year of training for each resident was then averaged with his or her co-residents, and the 5 resident classes were plotted over time (Figure 2). Remarkably, this approach also revealed inconsistent average score trends over time; 3 resident classes showed a statistically significant increase (p=0.001, 0.003, 0.006), 1 class showed a statistically significant decrease (p=0.008), and 1 class showed no significant change (p=0.16; one-way ANOVA with repeated measures). The varying trajectories of the clinical scores over time suggest that the evaluation tool was not consistently capturing improvement in clinical skill over time spent in residency, as all residents were deemed ready for graduation by the end of PGY4 via global evaluations. Also, the overall average mean OCEX score of 2.5 (± 0.4) in the first year of ophthalmology residency (PGY2) was higher than would be anticipated for a new resident given the grading scale of 1-4, suggesting an early tendency toward grade inflation and dampening the ability of the evaluation tool to capture real improvements in skill level later in training. In fact, our residents felt that specific evaluators were “easier” or “harder” graders using this tool.

In order to identify whether certain evaluators were more prone to grade inflation, we examined mean OCEX scores separated by evaluator. Although a total of 25 faculty evaluators submitted OCEX evaluations, only 4 faculty had submitted at least one evaluation for any resident in each of the 3 postgraduate year levels. Error bars depict standard deviation.

In our internal survey, some of the faculty expressed that they did not feel the OCEX was useful for resident education. To assess whether other ophthalmology residency programs were using OCEX, how useful it was for them, and potentially expand our dataset, we conducted a short email-based survey that asked faculty about their approach to evaluating residents. The survey was completed by 10 (67%) of the faculty who had filled out OCEX evaluations at one institution. The survey included multiple-choice questions that assessed whether the faculty assigned OCEX scores based on a relative versus absolute grading scale, i.e. relative to the expected level of skill for a resident at that level of training, versus an absolute scale comparing to the skill level expected of a fully trained, independent physician who is ready for graduation from residency (Figure 4A). Not surprisingly, the faculty provided a range of answers, demonstrating inconsistency of grading anchor interpretation. Evaluators were internally consistent with themselves whether they utilized a relative or absolute grading score, but the grading approach was not consistent across different evaluators (Figure 4B-C). This grading discrepancy helped to explain the erratic behavior of the mean OCEX data presented above, as different evaluators would assign low, medium, or high scores to the same first year resident with “average” clinical performance (Figure 4C). This occurred despite the availability of descriptive anchors for each point of the grading scale on the OCEX form, and despite quarterly faculty education on proper completion of the evaluation form.

Given the discrepancy in grading between evaluators, we conducted an anonymous survey to identify whether certain evaluators were more prone to grade inflation. We examined mean OCEX scores separated by evaluator. Although a total of 25 faculty evaluators submitted OCEX evaluations, only 4 faculty had submitted at least one evaluation for any resident in each of the 3 postgraduate year levels. Error bars depict standard deviation.

In order to identify whether certain evaluators were more prone to grade inflation, we examined mean OCEX scores separated by evaluator. Although a total of 25 faculty evaluators submitted OCEX evaluations, only 4 faculty had submitted at least one evaluation for any resident in each of the 3 postgraduate year levels. Error bars depict standard deviation.

In order to identify whether certain evaluators were more prone to grade inflation, we examined mean OCEX scores separated by evaluator. Although a total of 25 faculty evaluators submitted OCEX evaluations, only 4 faculty had submitted at least one evaluation for any resident in each of the 3 postgraduate year levels. Error bars depict standard deviation.
A Faculty Survey on OCEX Grading

1. When you assess resident skill level using the OCEX, do you compare a resident’s performance to (a) their PGY class, or (b) your own exam findings (i.e., the skill level of an independent practitioner)? (Answer: a or b)
2. The OCEX uses a grading scale of 1-4, where 4 = exceeds expectations and/or demonstrates technical facility in identifying subtle exam findings. What score would you typically give when clinical performance is average for PGY level for a first year resident? (Answer: 1, 2, 3, or 4)
3. What score would you typically give when clinical performance is average for PGY level for a third year resident? (Answer: 1, 2, 3, or 4)
4. If a first year resident and a third year resident made the same “rookie mistake” on clinical assessment of a patient, would you assign them a different numerical score on OCEX? (Answer: yes or no)
5. Do you feel the OCEX is useful for resident education? (Answer: yes or no) Optional: How is the OCEX useful or not useful?

B Grading Scale

C OCEX score assigned for “average” clinical performance

![Grading Scale](image)

![OCEX score assigned for “average” clinical performance](image)

Figure 4: Faculty survey.
(A) Internal faculty survey on OCEX grading behavior (B) to assess whether faculty used a relative or absolute grading scale. (C) Variable assessment by different evaluators for similar clinical performance, demonstrating inconsistency of grading anchor interpretation by different evaluators.

![Program director network survey on OCEX](image)

Figure 5: Program director network survey on OCEX.
This survey indicated that OCEX was used to evaluate trainees at approximately half of the ophthalmology residency programs.

poll an email listserv of ophthalmology residency program directors (EyePDnet, sponsored by the Association of University Professors of Ophthalmology). Approximately half (56 of 118) of the programs responded, of which slightly more than half (31) reported using OCEX. Programs used OCEX anywhere from once during residency to about 10 times per year of training, with an average of 3.3 (± 2.9) evaluations per year. Several included comments that centered on inconsistent faculty participation. Given the relative paucity of OCEX data across other programs, we were unable to expand our dataset further.

Discussion

Competency-based medical education for resident physicians is the way of the future, and work-based assessments of residents by clinical faculty are critical to the success of this approach [6]. The OCEX assessment tool has been validated for assessing clinical skills of ophthalmology residents for single “snapshot” patient encounters, but it was hitherto unclear whether this tool reliably captured the longitudinal development of resident skills over time. Demonstrating clinical progression over time would enable the tool to be used to help meet the ACGME mandate that such competency-based outcomes be documented. In our study, we instead found that OCEX scores did not consistently show the expected trend of improvement through the course of residency. Several residents even appeared to get worse over time! This result appeared to be due largely to varying interpretation of the grading scale anchors by evaluators. This non-standardization of grading is evident from our data in Figure 3, where the standard deviations of the scores across all graders did not even overlap, as well as the survey in which faculty openly admitted to using inconsistent grading scales (Figure 4). Thus a resident’s OCEX score appears to have depended more on which evaluator graded them than on the resident’s actual clinical skill level.

Our study raises important points about the challenges of rating performance. The adoption of a relative rather than an absolute grading scale, the latter being intended by the OCEX creators, by many of the evaluators impeded the use of our OCEX data to document resident skill progression over time. Our OCEX form contains grading scale criteria related to the meeting of expectations for “soft” skill competencies including professionalism, communication skills, and practice and systems-based learning. For the assessment of “technical” exam skills, discrete behavioral anchors were articulated describing facility with technique and whether exam findings were missed or identified. Yet we observed that faculty frequently evaluated residents in comparison with their class peers and/or relative to some expected level of skill associated with their year of training. The result is a “moving standard” effect that hinders efforts to document progression of proficiency. If a universal definition of what is expected of a resident is consistently applied, faculty should not be hesitant to assign a lower grade to the first-year resident who is expected to be less adept at the various technical aspects of the ophthalmic clinical exam. Furthermore, if expectations are delineated clearly, residents should not be disheartened by low marks early in training and instead see it as an opportunity for growth.

Our study has also revealed that faculty training on these assessments is insufficient to make them competent users of the OCEX tool. Our faculty are required to attend quarterly “faculty development” meetings at which OCEX anchors have been discussed. Despite these frequent reminders that absolute grading scale anchors should be used to gauge the accurate score of resident technical skill and that “meets expectations” should correlate with the ability of a graduating resident for soft skills, faculty admitted using self-determined standards that are not delineated in the OCEX assessment tool. Rigorous training such as that required of “expert raters” who administer assessments for research studies would be necessary for OCEX to accurately measure resident clinical skill. This standard is simply not feasible in the “real-world” clinical learning environment where assessment often ranks last among competing priorities for faculty’s limited time.

This retrospective study was limited by several factors. Notably, the OCEX does not formally instruct the grader on how to adjust for the degree of complexity of the patient’s problem, which could vary the difficulty of the resident’s examination. A number of OCEX forms were only partially completed, which may reflect time constraints or skipping of steps of the ophthalmic exam that were less relevant to the chief complaint. Additionally, the majority of OCEX forms were performed in the first year of residency. While presumably first year is when monitoring and feedback are most critical, the relatively fewer OCEX evaluations for senior residents may have increased individual marks early in training and instead see it as an opportunity for growth.

In addition to eradicating the “moving standard” effect and increasing faculty buy-in, promising strategies for improving clinical skill assessment may be inspired by innovative modalities used in surgery and business. For example, there are surgical evaluation tools that can distinguish between an ophthalmologist who has performed fewer than 200 cataract surgeries versus one who has performed over 1000 cataract surgeries [7]. These tools utilize objective data, such as decrease in phacoemulsification power and duration [8,9]. Other time-sensitive tools [10] to strengthen the accuracy of subjective competency assessments. Decreasing rates of surgical complications as residents accumulate surgical cases are another marker of trainee progress [10,11]. The relative value of different assessment tools may modeled by a formulaic relationship borrowed from the business world of change management: E = Q^A where E = effective change, Q = the technical quality of the solution, and A = the acceptance by the people involved [12]. Despite the high quality of OCEX with its demonstrated validity and reliability, its effectiveness as an assessment tool is hampered by relatively lower acceptance, which may be due to many reasons such as the length of time commitment, lack of recompense, grade inflation, and moving target scores due to relative and varying expectations. Indeed, this relationship may better explain the failure of mean OCEX scores to demonstrate discriminative validity, i.e. the ability to distinguish 1-3 years of residency training. Improving this acceptance factor will be vital before this tool can be used to show significant development in clinical proficiency. One way to appeal to graders is to simplify the evaluation process. For instance, wet laboratory microsurgical skill assessment tools list a series of tasks that the grader marks simply done correctly or not at all [13]. Over time, certain ophthalmic skillsets such as retinoscopy or manifest refraction may be rated as yes or no depending on the resident’s accuracy. Other aspects of OCEX may be amenable to short lists of structured, objective criteria or relevant examples of behavior corresponding to each number on the rating scale. Such specific guidance for grading will be essential as there is no gold standard for what defines a “good” level of skill. Programs and raters should firmly eliminate relative “moving standard” expectations in favor of absolute “graduation-ready” expectations that are explicitly defined.

Perhaps one inherent limiting feature of the standardized OCEX form is the four-point rating scale, which is coarse compared to the nine-point rating scale used in internal medicine’s mini-CEX [1,14]. Fine scales with more points tend to be more accurate than coarse scales for rating trainee behavior relative to scripted competence levels [15,16]. Thus, increasing the granularity of the rating scale may be a way to improve OCEX sensitivity and achieve discriminative validity. In addition, faculty can be routinely encouraged to make use of the full score range on OCEX. On the other hand, subjective experience with nine-point rating scales suggests that the full score range is rarely utilized and tendency is to skew toward higher scores even on arguably “average” performance [17].

We are not suggesting to eliminate or even to radically transform OCEX and other similar clinical skill evaluation tools. OCEX does a number of things well, such as facilitating multiple opportunities for formative feedback from faculty to residents and earlier identification of underperformers. Insightful, qualitative comments on OCEX forms

can help program directors to understand their residents’ strengths and weaknesses better for summative reviews. In general, use of clinical evaluation exercises promote the mission of the ACGME to ensure competency in patient care, but as seen in this study, these exercises may not be robust enough to provide meaningful longitudinal measures of performance. Instead, we argue that the benefits of OCEX should be maximized by improving faculty buy-in and refining the tool with more concrete metrics and expectations. Interestingly, the recently established Ophthalmology Milestones assess much more generic clinical skills (such as interview, exam, diagnostic procedures, and diagnosis) [18] than the skill-specific OCEX, so perhaps the OCEX will not be needed to monitor progression over time. Moreover, the Milestone levels of 1-5 feature an additional point on the rating scale as well as specific narrative anchors for each competency examined, both of which will foster improved precision in monitoring of longitudinal performance.

We report our experience with a clinical evaluation tool in hopes that it will aid other programs in implementing and innovating tools for longitudinal assessment. Future work is needed to verify our initial findings and to identify and test refinements in OCEX and other clinical competency measures. Armed with more standardized metrics and expectations, faculty buy-in should hopefully grow along with the correlated efficacy of the evaluation tool. Otherwise, the time and effort spent on these evaluations may be better applied toward other more efficacious assessment tools.

**Acknowledgements**

The authors thank Mae Gordon and Julia Huecker for their help in statistical analysis.

**Funding**

This work was supported by awards to the Department of Ophthalmology and Visual Sciences at Washington University from a Research to Prevent Blindness, Inc, unrestricted grant (New York, NY), the NIH Vision Core Grant P30 EY 0268 (Bethesda, MD). The funding organizations had no role in the design or conduct of this research.

**Ethical Approval**

An institutional review board (IRB) exemption was obtained for this medical education study (Washington University IRB# 200702096; University of Missouri IRB# 2008053).

**Previous Presentations**

An oral presentation describing preliminary data was given at the Educating the Educators meeting of the Association of University Professors of Ophthalmology, January 2016, Fort Lauderdale, FL.

**References**

**Supplemental Figure 1:** Modified Ophthalmic Clinical Evaluation Exercise (OCEX) form that was completed by faculty evaluators in this study. Each skill competency is individually graded on a rating scale ranging from 1-4, correlating to performance that is below expectations (score of 1), meets some expectations (score of 2), meets all expectations (score of 3), or exceeds expectations (score of 4), respectively.