Predicting Performance on Clerkship Examinations and USMLE Step 1: What Is the Value of Open-Ended Question Examination?

Christy K. Boscardin, PhD, Gillian Earnest, MS, and Karen E. Hauer, MD, PhD

Abstract

Purpose
Despite advances in learning sciences that highlight the efficacy of elaborative interrogation, in which students explain and elaborate on concepts in their own words, assessment techniques in medical education have commonly employed multiple-choice questions (MCQs). Educators’ reluctance to consider alternatives such as open-ended questions (OEQs) stems from practical advantages of MCQs and the lack of empirical data on the predictability of OEQs for performance on other high-stakes assessments. In this study, the authors compared the predictive value of preclerkship assessments using OEQs for the outcomes of clerkship examinations and United States Medical Licensing Examination (USMLE) Step 1.

Method
The authors compared outcomes of 2 assessment formats using multiyear performance data (2015 and 2016 cohorts) on preclerkship MCQ versus OEQ examinations for predicting students’ subsequent performance on 6 clerkship examinations and USMLE Step 1. The authors conducted a regression analysis to compare the predictability of MCQs and OEQs by using clerkship exam scores and Step 1 scores as dependent variables and performance on MCQs and OEQs as predictors in the models.

Results
Regression models with OEQs were consistently higher for predicting clerkship exam (NBME shelf-exam) scores, except for one clerkship, compared with models using MCQs. For Step 1, R-square using MCQs was higher with 59% of the variance explained compared with 46% with OEQs, but the OEQ cohort scored significantly higher on Step 1.

Conclusions
OEQ examinations predict performance on subsequent high-stakes MCQ examinations. Given the predictive value and closer alignment with scientific principles of effective learning, OEQ examinations are an examination format worthy of consideration in preclerkship medical education programs.

With the proliferation of medical and scientific knowledge and advancements in information technology, application of medical knowledge and focus on conceptual understanding rather than recall or factual knowledge alone have become critical components of preparation for 21st-century physician practice. To address these needs, medical schools are seeking alternative approaches to prepare trainees to manage their own medical knowledge learning within the changing landscape of health care. This shift in medical education emphasizes the importance of conceptual understanding (deeper learning) and application of knowledge for long-term retention rather than just immediate recall of facts.

Over the past several decades, study of the science of learning has added to understanding of effective teaching, learning, and assessment strategies and more recently popularized these concepts in the book Make It Stick: The Science of Successful Learning. Two major assessment strategies that promote long-term knowledge retention include elaborative interrogation (asking and explaining why and how things work) and retrieval practice (bringing learned information to mind from long-term memory). The strategy of elaborative interrogation entails the learner asking and discussing “how” and “why” questions about important concepts. This approach has shown to enhance learning across multiple domains for learners in various settings including elementary, middle and high school, and undergraduate education. The prevailing theoretical explanation for the efficacy of elaborative interrogation is that it activates schemata and knowledge structure for efficient integration of new information with prior knowledge. Teachers can use elaborative interrogation as part of their assessment strategy by emphasizing and valuing self-explanation and critical thinking. Another key way that assessment can facilitate learning is through retrieval practice. Decades of research on memory have shown that the act of retrieval, as occurs during test-taking, strengthens and improves retention.

Despite these advances in learning sciences, assessment techniques in medical education have commonly employed multiple-choice questions (MCQs) and have been slow to change or advance from this knowledge assessment format. MCQs have some very practical advantages over other formats including open-ended questions (OEQs) such as ease of grading, potential opportunity to test a broad number of concepts, and sense of objectivity in grading. Although some studies have also proposed that MCQs can assess higher-order cognitive skills similar to OEQs, the controversy still remains whether the 2 formats are indeed measuring the same constructs and level of understanding. OEQs invite students to generate explanations using their own words and therefore help operationalize elaborative interrogation principles. OEQs have
advantages over MCQs because of the opportunity to pose “why” and “how” questions that require explanation of concepts and relevance to clinical scenarios, drawing on higher-order thinking and problem-solving skills. Additionally, studies suggest that OEQs have positive impact on anticipatory learning and conceptual understanding, while MCQs tend to favor memorization and recall.

A major concern that may explain the limited adoption of OEQs in medical education is the lack of empirical data on the usefulness of OEQs to predict performance on other high-stakes assessments which are often MCQs. The reduction in sampling of the domain coverage and thus internal consistency (reliability) could contribute to overall lower predictability to future performance. Specifically in medical education, concerns by students and faculty over missing the opportunity for students to practice “simulated board exam” MCQs during preclerkship courses and prepare for the format of United States Medical Licensing Examination (USMLE) Step 1 may deter schools from considering implementation of OEQs. The reduced number of domains or topics that can be targeted during the exam is one of the major drawbacks of OEQs, but it can also drive the exam to be more focused on essential test objectives and encourage questions that assess multiple aspects of a concept or case. A comparison of the predictability of MCQs and OEQs on other exam performance will provide additional information for consideration in assessment format design and selection.

As part of our programmatic assessment, our institution implemented an assessment framework that reflects learning science principles. Specifically, we implemented OEQs, also referred to as constructed-response questions, to emphasize and facilitate self-explanation, to encourage students to consider why and how questions, and to generate practice opportunities for elaborative interrogation. We designed this study to compare the predictive value of preclerkship OEQs with MCQs used in prior curriculum on the outcomes of MCQ clerkship exams and USMLE Step 1 by harnessing the data from before and after changing our preclerkship assessment format from MCQs to OEQs.

**Method**

This is a comparison study of 2 assessment formats using multiyear data to examine student performance on preclerkship and clerkship examinations and Step 1. The institutional review board of our institution approved the study as exempt.

**Setting**

Our urban public institution implemented a new assessment format within a new curriculum in fall 2016. Before 2016, in the legacy curriculum (2015 cohort), the preclerkship curriculum contained 7 integrated courses over 2 years. The 2015 cohort took Step 1 before the start of their clerkships. With the entering class of 2016 (new curriculum), the preclerkship curriculum comprised 10 integrated 3- to 5-week blocks over 18 months. The entering 2016 cohort took Step 1 after completing their core clerkships, a schedule that created a lag time between completion of their preclerkship courses and Step 1.

**Sample**

We used available performance data from all medical students in the entering class of 2015 (legacy curriculum) and 2016 (new curriculum) to conduct a direct comparison of performance on preclerkship MCQ and OEQ summative examinations with clerkship performance and Step 1 performance.

**Measures**

**Preclerkship examinations for legacy curriculum.** For the entering class of 2015, all summative examinations in preclerkship courses consisted of MCQs. Questions were authored by faculty and represented the test item format of USMLE national board assessment questions. Faculty responsible for the course content received training on item writing. Items for the exam were reviewed by the course directors and then modified as needed based on the item statistics report following the administration. We used percentage of correct responses on each exam for the analysis.

**Preclerkship examinations for new curriculum.** With the entering class of 2016, all summative examinations in our preclerkship courses used an OEQ format. Questions present a clinical history scenario and students construct a short response (typically 1 or 2 paragraphs) to questions (ranging from 9 to 16 questions per exam) using their own words. There are 13 summative OEQ examinations in the 10 courses. Student responses are graded using a rubric that categorizes each response into “meets expectations,” “borderline,” or “does not meet expectations.” In this study, we used the percentage of questions rated as “meets expectations” for each exam to represent student performance on OEQs. Students have the opportunity to review their performance compared with the model answers during a 1-hour in-person exam viewing session shortly after the completion of the grading.

Additionally, the curriculum embedded weekly quizzes in OEQ (2–3 questions) and MCQ (20–25 questions) formats as part of the formative assessment to guide learning (no grading) and to trigger retrieval practice. These weekly quizzes are not graded/collated and used only to provide opportunities for self-assessment.

**Clerkship examinations.** For both student cohorts, we used clerkship examination scores from core 6 clerkships that administered scored exams: Medicine, Neurology, Obstetrics–Gynecology, Pediatrics, Psychiatry, and Surgery. Five clerkships administered NBME subject examinations. The Pediatrics examination is a national case-based exam (Aquifer). All clerkship examinations are MCQs.

**USMLE Step 1 scores.** We used available Step 1 scores from both cohorts including all Step 1 scores reported by July 2019.

**Analysis**

We conducted descriptive statistics on the performances on all the exams. Next, we compared student performance on clerkship exams and Step 1 scores using *t* tests between 2015 and 2016 cohort with the Bonferroni correction to account for multiple comparisons. We conducted a regression analysis to compare the predictability of MCQs and OEQs by using clerkship exam scores and Step 1 scores as dependent variables and performance on MCQs and OEQs as predictors in the models. We compared the R-squares which capture the amount of variance explained by the models using MCQs or OEQs as (the preclerkship assessment performance) predictors of...
the clerkship and Step 1 scores and the correlations among the variables.

Results

Demographics

We report on the findings from 258 students with complete data for preclerkship, clerkship, and Step 1 (131 in the 2015 cohort and 127 in the 2016 cohort) out of 306 total eligible students. Nineteen students (5 from 2015 and 14 from 2016) with missing Step 1 scores and 29 students (15 from 2015 and 14 from 2016) with one or more missing clerkship exams were excluded from the study. The proportion of students self-identifying as under-represented in medicine (African American, Latinx, Native American, Alaska Native, Native Hawaiian, or Pacific Islander) was slightly lower in 2015 (31%) compared with 2016 (36%). The breakdown of gender was similar between 2015 and 2016 cohorts (49% and 50% female students, respectively).

MCQ exam scores

The average score on the 7 end-of-course examinations ranged from 85% (standard deviation [SD] = 6) to 87% (SD = 5) correct for the 2015 cohort.

OEQ exam scores

The 13 end-of-course examination score averages ranged from 75% (SD = 15) to 86% (SD = 12) of the exam items rated as “meets expectations” for the 2016 cohort.

Clerkship exam scores

The 2015 cohort scored significantly higher on 4 of the 6 clerkship examinations compared with the 2016 cohort, with percentiles ranging from 74% (SD = 8) to 83% (SD = 6) for 2015 and 69% (SD = 10) to 81% (SD = 7) for 2016 (see Table 1).

Step 1 performance

As shown in Table 2, the 2016 cohort scored significantly higher on Step 1 (X = 240, SD = 16) compared with the 2015 cohort (X = 234, SD = 18; P = .006).

Predictability comparison

As shown in Table 3, R-squares (percent of variance explained) for the model with OEQs were consistently higher for predicting clerkship exam scores except for the Medicine clerkship (no difference) compared with models using MCQs. The percent of variance explained ranged from 37% to 46% with MCQs and 44% to 52% with OEQs. For Step 1 scores, R-square using MCQs as predictors was higher with 59% of the variance explained compared with 46% with OEQs (Table 2).

Discussion

Despite the potential advantages of OEQs over MCQs for promoting desirable learning behaviors, a minority of schools have adopted OEQs as a core assessment strategy due to concerns about the predictability and potential impact on clerkship and Step 1 examination performance, in addition to logistic concerns. In this study, we found that the predictability of preclerkship exam performance on clerkship exam performance was significantly higher with OEQs compared with MCQs. Step 1 scores were also higher for the cohort with OEQs compared to cohort with MCQs. Not surprisingly, clerkship exam scores were slightly lower for the 2016 (OEQ) cohort, given that their clerkships started about 6 months earlier and they had not yet had a dedicated study period preparing for Step 1. Additionally, some clerkship exam mean scores were lower for the cohort with OEQs; however, the predictability of OEQs, compared with MCQs, was overall consistently higher for clerkship exams.

Similar to previous studies examining the impact of moving Step 1 to after the core clerkships, we found significant increases in Step 1 scores with the new curriculum cohort.28,29 This increase in the scores may reflect that students had more time to consolidate their knowledge and practice with MCQ-type exams during clerkships with the new Step 1 exam schedule. Students may also benefit from the opportunity to familiarize themselves with the testing format of a board exam through experience with NBME clerkship exams during clerkships. However, despite the greater lag time between the preclerkship curriculum and Step 1, we found that substantial variance in Step 1 performance was accounted by their OEQ performance. Despite the hypothesis that MCQs during preclinical curricula may be uniquely suited to prepare students for Step 1, our findings suggest that the practice effect from the particular test format may not be as important for Step

Table 1

<table>
<thead>
<tr>
<th>Clerkship</th>
<th>2015 cohort (MCQ), mean (SD)</th>
<th>2016 cohort (OEQ), mean (SD)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine</td>
<td>75 (8)</td>
<td>69 (10)</td>
<td>&lt; .001*</td>
</tr>
<tr>
<td>Neurology</td>
<td>78 (7)</td>
<td>76 (9)</td>
<td>.010</td>
</tr>
<tr>
<td>OB/GYN</td>
<td>78 (7)</td>
<td>77 (7)</td>
<td>.184</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>83 (6)</td>
<td>81 (7)</td>
<td>.019</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>80 (6)</td>
<td>79 (7)</td>
<td>.050</td>
</tr>
<tr>
<td>Surgery</td>
<td>74 (8)</td>
<td>71 (9)</td>
<td>.001*</td>
</tr>
</tbody>
</table>

Abbreviations: MCQ, multiple-choice question; OEQ, open-ended question; SD, standard deviation; OB/Gyn, obstetrics and gynecology.
*Statistically significant after adjusting for multiple comparisons with the Bonferroni correction (P < .008).

Table 2

<table>
<thead>
<tr>
<th>Year</th>
<th>Step 1, mean (SD)</th>
<th>t test value</th>
<th>R-square</th>
<th>Average correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015 (MCQ)</td>
<td>234 (18)</td>
<td>.006</td>
<td>59%</td>
<td>0.6</td>
</tr>
<tr>
<td>2016 (OEQ)</td>
<td>240 (16)</td>
<td>.46</td>
<td>46%</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Abbreviations: USMLE, United States Medical Licensing Examination; MCQ, multiple-choice question; OEQ, open-ended question; SD, standard deviation.
need a programmatic approach to early identification, struggling or at-risk students who may activate preventative measures for early in the curriculum. Preclerkship for identification of struggling learners early performance indicators are available is the demonstration that trustworthy students to maximize the benefits of additional insight.

Our findings underscore the opportunity for early identification of struggling students to maximize the benefits of remediation and additional support. One positive benefit of this predictive study is the demonstration that trustworthy early performance indicators are available for identification of struggling learners early in the curriculum. Preclerkship performance data can be used as a trigger for providing additional resources and activating preventative measures for struggling or at-risk students who may underperform in clerkships. In addition to early identification, struggling students need a programmatic approach to remediation for a successful outcome. As discussed by other authors, design of a remediation program should represent and incorporate the principles of programmatic assessment with emphasis on each learners’ development through self-monitoring and effective feedback. At our institution, we have introduced a student performance dashboard to facilitate self-regulated learning with the expectation that learners will monitor their own performance and set individual learning goals; this infrastructure enables both identification of students needing additional resources and early support for them. Detailed longitudinal views of examination performance to showcase the longitudinal data that predict future performance, as shown in this study, enable both learning planning and individualized coaching to maximize the utility of exam data from preclerkship courses. Coaching can help normalize the concept of continuous improvement, thereby empowering students to review and interpret their own performance data and to seek additional resources when necessary.

Our study set out to address the concerns around the potential impact of a different testing format on subsequent Step 1 performance. For most of the 20th century, the emphasis on predictive validity arguments favored MCQs over OEQs. Our empirical study demonstrates that OEQs can be as predictive as MCQs of later performance on high-stakes MCQ exams of medical knowledge, but more importantly may facilitate different learning approaches that may be more durable. Despite this positive finding, we would advocate for testing format selection based on theoretically and constructively driven approach. Given the changes in our curriculum including timing of Step 1, the increase in Step 1 scores is encouraging but may be attributable to multiple factors.

Using Messick’s unitary view of validity, we advocate for evaluating the quality of the exam (format) beyond just predictive validity by incorporating the review of intended and unintended consequences of the assessment format on learning. The change in our assessment format was driven by our desire to facilitate more favorable learning behaviors and outcomes. OEQs require additional time for construction and grading compared with MCQs. Despite this additional time commitment, most of our faculty find OEQs to be more informative and align with the learning objectives of the course. The key next step in our evaluation of the OEQs will focus on the impact of these assessments on student learning behaviors and strategies and long-term outcomes in clinical reasoning as well as clinical performance. A shift in students’ learning behaviors toward strategies that focus more on conceptual understanding than memorization would provide additional evidence of desirable impact of our new assessment format.

Our study has several limitations. First, the findings are based on a single institution with data from 2 cohorts of students. Additionally, more contextual information on students’ exam preparation strategies for both clerkship exams and Step 1 would have provided more insights into our findings. Also, the data were collected during a curricular transition which may have impacted the performance comparison between the 2 cohorts. Lastly, longitudinal study of the impact of the assessment format on long-term retention in clinical practice would provide more detailed information on the utility and further validity evidence of OEQs on outcomes.

Several competing factors and resources contribute to decision making regarding assessment strategies and format implementation in a curriculum. Our findings showed that despite some of the logistical challenges of OEQs compared with MCQs, the predictive value and closer alignment with the scientific principles of effective learning of OEQs demonstrate their advantages over MCQs for further consideration in assessment selection and design.

Acknowledgments: The authors wish to thank Patricia O’Sullivan, EdD, for her guidance and feedback on earlier drafts of this article.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: The institutional review board of the authors’ institution approved the study as exempt.

C.K. Boscardin is associate professor, Department of Medicine, University of California, San Francisco, San Francisco, California; ORCID: https://orcid.org/0000-0002-9070-8859.

G. Earnest is research analyst, Center for Faculty Educators, University of California, San Francisco, San Francisco, California.

K.E. Hauer is associate dean for assessment and professor, Department of Medicine, University of California, San Francisco, San Francisco, California; ORCID: https://orcid.org/0000-0002-8812-4045.

References
5 Dunlosky J, Rawson KA, Marsh EJ, Nathan MI, Willingham DT. Improving students’


21 Livingston SA. Constructed-response test questions: Why we use them; how we score them. Educ Test Serv. 2009;11:1–8.


