Evaluating Student Clerkship Performance Using Multiple Assessment Components

Oladimeji Oki, MD | Zoon Naqvi, MBBS, EdM, MHPE | William Jordan, MD, MPH | Conair Guilliamenes, MD | Heather Archer-Dyer, MPH, CHES | Maria Teresa Santos, MD

PRiMER. 2024;8:25.

Published: 4/23/2024 | DOI: 10.22454/PRiMER.2024.160111

Abstract

Introduction: Family medicine clerkships utilize a broad set of objectives. The scope of these objectives cannot be measured by one assessment alone. Using multiple assessments aimed at measuring different objectives may provide more holistic evaluation of students. A further concern is to ensure longitudinal accuracy of assessments. In this study, we sought to better understand the relevance and validity of different assessment tools used in family medicine clerkships.

Methods: We retrospectively correlated family medicine clerkship students' scores across different assessments to evaluate the strengths of the correlations, between the different assessment tools. We defined $\rho < 0.3$ as weak, $\rho > 0.3$ to $\rho < 0.5$ as moderate, and $\rho > 0.5$ as high correlation.

Results: We compared individual assessment scores for 267 students for analysis. The correlation of the clinical evaluation was 0.165 ($P < .01$); with case-based short-answer questions it was 0.153 ($P < .01$); and with objective structured clinical examinations it was -0.246 ($P < 0.01$).

Conclusion: Overall low levels of correlations between our assessments are expected, as they are each designed to measure different objectives. The relatively higher correlation between component scores supports convergent validity while correlations closer to zero suggest discriminant validity. Unexpectedly, comparing the multiple-choice questions and objective, structured clinical encounter (OSCE) assessments, we found higher correlation, although we believe these should measure disparate objectives. We replaced our in-house multiple-choice questions with a nationally-standardized exam and preliminary analysis shows the expected weaker correlation with the OSCE assessment, suggesting periodic correlations between assessments may be useful.

Introduction

Family medicine (FM) clerkships require students to achieve a broad set of objectives including clinical knowledge, verbal and written communication, physical examination, and analytic skills. These objectives can be difficult to measure with a solitary assessment. Therefore, many FM clerkships employ a combination of assessment strategies to evaluate medical students’ objective achievement. However, few clerkships have formally evaluated the quality, relevance, and validity of the multiple assessments that determine the final grade to ensure they accurately depict students' performance. We developed a conceptual framework to assess
the theoretical and empirical relationship between the different assessment tools used in our clerkship.\textsuperscript{4,5} Here we present the results of the correlation analyses between multiple assessments to determine the relevance and validity of each assessment component and highlight recommended changes.\textsuperscript{1-3}

**Conceptual Framework**

The FM clerkship has historically used multiple components to evaluate student performance. Table 1 demonstrates how we developed a framework to map the relationship between the assessment tools and clerkship objectives.\textsuperscript{4,5} When writing clerkship objectives, the clerkship team aimed to keep the objectives clear, concise, measurable, and closely aligned to the specific goals and learning outcomes of our institution's overall educational program objectives. The framework was developed, reviewed, and refined through data analysis of scores and educational literature over several years by multiple members of the clerkship team.\textsuperscript{4, 5} The final grade in the required FM clerkship includes the following components:

- Clinical evaluation (CE),
- Multiple-choice exam (MCQs),
- Case-based short-answer questions (CBSA),
- Objective structured clinical encounters (OSCEs), and
- A community project (CP) handoff and advisor score.

These scores evaluate students' performance in clinical knowledge, procedural and documentation skills, clinical reasoning, interpersonal skills, teamwork, and implicit and explicit attitudes. None of our assessment tools alone can assess the achievement of all objectives.\textsuperscript{6} Table 1 maps our assessments to the degree of expected measurement of our objectives.

**Validity and Reliability Measures**

Constructing multiple, validity-related hypotheses to assess whether comparative assessment tool scores reflect certain abilities is not always straightforward.\textsuperscript{2,7-13} We used Pearson's correlation coefficient ($\rho$) to study the relationship between the different tools (scale: 1 to -1), allowing a measure of the similarity of multiple assessment scores.\textsuperscript{2,10,12} For this review, we determined that $\rho<0.3$ represented weak correlation, $\rho\geq0.3$ to $\rho<0.5$ moderate correlation, and $\rho\geq0.5$ represented high correlation.\textsuperscript{13} While weaker levels of correlation overall are expected, as no two assessments are intended to measure the same objectives, we also expect tools that have significant overlap in intended assessed objectives (eg, OSCEs and CBSA), to show relatively higher levels of correlation suggesting convergent validity (CV, here defined as $\rho>0.2$). Likewise, those tools measuring disparate objectives, such as the CE and the CP, will support discriminant validity (DV), or $\rho$ closer to 0. Table 2 reflects our group's hypotheses on how different tools relate to each other in terms of validity, based on the intended measured objectives.

**Methods**

**Study Design**

We performed a retrospective correlation analysis of a pre-existing database containing medical students' scores for component assessments during their FM rotations. We compared all students from 2 academic years (2018-2020) with the same preclinical training. Due to the COVID-19 pandemic, OSCEs were discontinued, and CE scores were modified to pass/fail for the last 3 of the 12 rotations for 2019-2020. Therefore, for balance, we included all students from the first 9 rotations of the 2 academic years in the analysis. Albert Einstein College of Medicine deemed our study exempt from approval (IRB #: 2019-10288).

The required 4-week FM clerkship takes place in the third year of medical school training. The final grade
(honors, high pass, pass, low pass, or fail) is determined using criterion-based cutoffs of the aggregate scores.

**Data Analysis**

We entered assessment scores into a correlation matrix representing correlation for the combined 2 years using SPSS 24.0 software. We then compared the correlation analysis to the hypothesized level of correlation (convergent vs divergent validity) based on the predefined intended assessment measurement of objectives.

**Results**

Table 3 shows a detailed correlation matrix using all components of the evaluation to examine correlation between components. We included data for 267 students in 2 academic years (2018-2019 and 2019-2020). Table 3 also shows the correlations between the component scores. All correlations were overall weak to moderate, varying from a maximum of 0.36 to -0.044.

**Discussion**

The overall weak-to-moderate correlations among our assessment tools were expected and support the use of multiple assessments in measuring student performance during the FM clerkship. Having multiple assessments allows us to evaluate achievement of multiple, disparate clerkship objectives leading to a more holistic assessment of student performance. Our *a priori* assumption was that the correlation between components that assess more overlapping objectives (eg, OSCEs and CBSA) would support CV. Interestingly, some results did not support our assumptions. For example, the correlation of the MCQ scores (intended to measure clinical recall and limited clinical reasoning) with OSCEs (measuring communication), although moderate, are among the strongest correlations in our data ($\rho = 0.325$, Table 3) and does not support the expected DV. This suggests that these components may be assessing unintended objectives. For example, our MCQs may be assessing verbal/reading objectives as well, or our OSCEs may be measuring more clinical recall/reasoning than intended. Our MCQs did not meet the expected hypothesis with our CBSA nor our CP handoff score. As expected, the correlations of the CP and the community advisor scores with the other scores were closer to 0, suggesting stronger DV.

**Limitations**

Individual bias is a potential confounder in measuring assessed objectives, and this paper does not explicitly address such bias. Regardless of how thorough a grading rubric is, individual graders may still interpret the rubric differently. For example, clinical evaluation by preceptors has historically been difficult to standardize and offers ample opportunity for bias/unintended objective evaluation. This type of evaluation needs further standardization by developing a more user-friendly rubric, and continued faculty development to minimize the potential for nonmeasured objectives to be included in the assessment.

A second confounder is the unequal weighting of assessments. This inequality could potentially influence students to put more effort into more highly weighted assessments over others, introducing further variables into our correlation scores. As our weighting system was designed so that students needed to do well in all components to achieve honors for the clerkship, we do not believe this is a significant confounder. However, assessment weighting is a topic that should be studied further.

**Conclusions**

Evaluating assessments in a comparative format allowed us to identify gaps in the validity of our assessments. As our assessments were created to measure student knowledge, skills, and behaviors in relation to our course...
objectives, we aim for them to accurately assess said objectives. We use multiple assessments to measure a disparate set of objectives that cannot be measured with one assessment alone. Ideally, assessment scores should have weak correlation with one another; otherwise the multiple assessments may be unnecessary and redundant. Correlating assessment scores allowed our clerkship team to better understand the validity of our assessments by noting if the assessments are measuring the intended objectives. Based on our results, we replaced our in-house MCQs with a nationally standardized exam to assess clinical knowledge more reliably. Preliminary analysis indicates weaker correlation of our new MCQs with OSCEs and CE, a scenario better aligned to what we initially hypothesized. Regularly assessing clerkship-grading components in this manner provides an opportunity to contribute to validation of the measures and ensure they are properly assessing the course objectives.

Tables and Figures

Table 1. Mapping Objectives to Assessment Components

<table>
<thead>
<tr>
<th>Objective Description</th>
<th>Clinical evaluation</th>
<th>MCQ exam</th>
<th>CBSA questions</th>
<th>OSCEs</th>
<th>CP handoff score</th>
<th>Community site advisor assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obtain a relevant patient history</td>
<td>+++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conduct a physical examination</td>
<td>+++</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demonstrate sound clinical reasoning</td>
<td>++</td>
<td>+</td>
<td>+++</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formulate a patient-centered management plan</td>
<td>++</td>
<td>+</td>
<td>+++</td>
<td>++</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop an evidenced-based health maintenance plan</td>
<td>+++</td>
<td></td>
<td>+</td>
<td>+++</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demonstrate written and verbal communication skills</td>
<td>++</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>Document accurate and relevant information</td>
<td>++</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implement relevant quality improvement (QI) project</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Discuss the role of population-level determinants on the QI initiative identified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Recommend available community assets and resources to improve the health status of the target population</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Demonstrate team skills</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Weighting in final clerkship grade (%)</td>
<td>45</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

Abbreviations: MCQ, multiple choice question exam; CBSA, case-based short-answer questions; OSCE, objective structured clinical encounters; CP, community project.
Clerkship objectives were created based on mapping our institution’s Educational Program Objectives. The clerkship team then reviewed each assessment for the projected degree of skill measurement for each objective, with + being partially assessed (<25% of assessment measures this objective), ++ moderately assesses (25%-50%), and +++ being significantly assessed (>50%).
1. Core clinical site directors aggregated preceptor scores based on student clinical performance.
2. Two versions of an in house generated 50 MCQ exam adapted from a validated FM question bank.
3. Two versions of set of two case presentation based short answer prompts assessing student ability to generate a history, physical exam, differential diagnoses, assessment, and plan, associated with a structured checklist/rubric (the MCQs and CBSA are part of a single 2-hour final exam).
4. Six standardized history taking only patient stations and six stations of associated long menu questions based on the preceding patient station.
5. Group written assignment based on a student driven quality improvement project assessed with a structured rubric.
6. Structured summary of the team activities assessed with a structured rubric/scale assessed by a community site advisor.
Table 2. Hypotheses and Analyses Using Comparison of Assessment Components

<table>
<thead>
<tr>
<th>Assessment component</th>
<th>Learning objective(s)</th>
<th>Clinical evaluation</th>
<th>MCQ exam</th>
<th>CBSA</th>
<th>OSCEs</th>
<th>CP handoff score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical evaluation</td>
<td>Obtain a relevant history, conduct a PE, and clinical reasoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCQ exam (50 items)</td>
<td>FM knowledge/content (objective and standardized)</td>
<td>Discriminant validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBSA questions</td>
<td>Clinical reasoning skills and knowledge</td>
<td>Convergent validity</td>
<td>Discriminant validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSCEs (12 stations)</td>
<td>Contextual communication skills and follow-up open-book paper cases</td>
<td>Convergent validity</td>
<td>Discriminant validity</td>
<td>Convergent validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP handoff score</td>
<td>Written communication skills, including logical organization</td>
<td>Discriminant validity</td>
<td>Discriminant validity</td>
<td>Discriminant validity</td>
<td>Discriminant validity</td>
<td></td>
</tr>
<tr>
<td>Community site advisor score</td>
<td>Demonstrate team skills in a learning and service context</td>
<td>Discriminant validity</td>
<td>Discriminant validity</td>
<td>Discriminant validity</td>
<td>Discriminant validity</td>
<td>Convergent validity</td>
</tr>
</tbody>
</table>

Hypotheses were generated based on the degree of measurement overlap of our objectives by each of our assessments. Assessments that had more overlap in objectives measured were expected to show convergent validity while those measuring different objectives were expected to show discriminant validity.

Abbreviations: MCQ, multiple choice question exam; CBSA, case-based short-answer questions; OSCE, objective structured clinical encounters; CP, community project; PE, physical exam; FM, family medicine.

Table 3. Correlations Between Components, Combined Years 1 and 2 (AY1 n=133, AY2; n=134; N=267)

<table>
<thead>
<tr>
<th></th>
<th>CE</th>
<th>MCQ</th>
<th>CBSA Questions</th>
<th>OSCEs</th>
<th>CP handoff score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE (Objectives 1-7)</td>
<td>1</td>
<td>0.165**</td>
<td>0.362**</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MCQ Exam (Objectives 3-5)</td>
<td>0.153*</td>
<td>0.246**</td>
<td>0.325**</td>
<td>0.268**</td>
<td>1</td>
</tr>
<tr>
<td>CBSA Questions (Objectives 1-4, 6)</td>
<td>0.223**</td>
<td>0.235**</td>
<td>1</td>
<td>-0.019</td>
<td>-0.044</td>
</tr>
<tr>
<td>OSCEs (Objectives 1, 4-6)</td>
<td>0.235**</td>
<td>0.246**</td>
<td>0.325**</td>
<td>1</td>
<td>0.102</td>
</tr>
<tr>
<td>CP handoff score (Objectives 6-10)</td>
<td>0.235**</td>
<td>0.246**</td>
<td>0.325**</td>
<td>1</td>
<td>0.102</td>
</tr>
<tr>
<td>Community Site Advisor Assessment (Objectives 6, 11)</td>
<td>0.235**</td>
<td>0.246**</td>
<td>0.325**</td>
<td>1</td>
<td>0.102</td>
</tr>
</tbody>
</table>

Abbreviations: CE, clinical evaluation; MCQ, multiple choice question exam; CBSA, case-based short-answer questions; OSCE, objective structured clinical encounters; CP, community project.

* Correlation is significant at the 0.05 level (2-tailed)
** Correlation is significant at the 0.01 level (2-tailed)

Due to the pandemic, OSCEs were dropped and CEs were modified for the last 3/12 rotations for AY2019-20. Therefore, for balance, we included all students of the first 9 rotations for the 2 academic years in the analysis. Shaded cells are areas where our hypotheses were not supported by the correlations.

Acknowledgments

Presentations: Data from this article were presented at the STFM conference on Medical Student Education in Austin, Texas on February 3, 2018, as well as at the Family Medicine Education Consortium conference in Rye, New York on November 10, 2018.

Corresponding Author

Oladimeji Oki, MD
References


Copyright © 2024 by the Society of Teachers of Family Medicine