PRiMER
peer-reviewed reports in medical education research
research brief

Evaluating Student Clerkship Performance Using Multiple Assessment Components

Oladimeji Oki, MD | Zoon Naqvi, MBBS, EdM, MHPE | William Jordan, MD, MPH | Conair Guilliames, 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.[1] 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.[1-3] We developed a conceptual framework to assess

the theoretical and empirical relationship between the different assessment tools used in our clerkship.[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.[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.[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.[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.[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.[2,7-13] We used Pearson's correlation coefficient (ρ) to study the relationship between the different tools (scale: 1 to -1), allowing a measure of the similarity of multiple assessment scores.[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.[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 ρ 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 (ρ= 0.325, Table 3) and does not support the expected DV.[6] 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 to 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

	Clinical evaluation[1]	MCQ exam[2]	CBSA questions[3]	OSCEs[4]	CP handoff score[5]	Community site advisor assessment[6]
Obtain a relevant patient history	+++		+	+		
Conduct a physical examination	+++		+			
Demonstrate sound clinical reasoning	++	+	+++			
Formulate a patient-centered management plan	++	+	+++	++		
Develop an evidenced-based health maintenanceplan	+++	+		+++		
Demonstrate written and verbal communication skills	++		++	+++	+++	++
Document accurate and relevant information	++				+++	
Implement relevant quality improvement (QI) project					+++	
Discuss the role of population-level determinants on the QI initiative identified					+++	
Recommend available community assets and resources to improve the health status of the target population					+++	
Demonstrate team skills	++					+++
Weighting in final clerkship grade (%)	45	15	15	10	10	5

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 assesses (<25% of assessment measures this objective), ++ moderately assesses (25%-50%), and +++ being significantly assesses (>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

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

	CE	MCQ	CBSA Questions	OSCEs	CP handoff score
CE (Objectives 1-7)	1				
MCQ Exam (Objectives 3-5)	0.165**	1			
CBSA Questions (Objectives 1-4, 6)	0.153*	0.362**	1		
OSCEs (Objectives 1, 4-6)	0.246**	0.325**	0.268**	1	
CP handoff score (Objectives 6-10)	0.082	0.223**	0.235**	0.102	1
Community Site Advisor Assessment (Objectives 6, 11)	0.016	-0.012	-0.008	-0.019	-0.044

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

Albert Einstein College of Medicine, Bronx, NY
ooki@montefiore.org

Author Affiliations

Oladimeji Oki, MD - Albert Einstein College of Medicine, Bronx, NY
Zoon Naqvi, MBBS, EdM, MHPE - Albert Einstein College of Medicine, Bronx, NY
William Jordan, MD, MPH - New York City Department of Health & Mental Hygiene, New York, NY
Conair Guilliames, MD - Albert Einstein College of Medicine, Bronx, NY
Heather Archer-Dyer, MPH, CHES - Albert Einstein College of Medicine, Bronx, NY
Maria Teresa Santos, MD - Albert Einstein College of Medicine, Bronx, NY

References

1. Shumway JM, Harden RM; Association for Medical Education in Europe. AMEE Guide No. 25: the assessment of learning outcomes for the competent and reflective physician. *Med Teach*. 2003;25(6):569-584. doi:10.1080/0142159032000151907
2. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830-837. doi:10.1046/j.1365-2923.2003.01594.x
3. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20(12):1159-1164. doi:10.1111/j.1525-1497.2005.0258.x
4. Lee M, Wimmers PF. Clinical competence understood through the construct validity of three clerkship assessments. *Med Educ*. 2011;45(8):849-857. doi:10.1111/j.1365-2923.2011.03995.x
5. McLaughlin K, Vitale G, Coderre S, Violato C, Wright B. Clerkship evaluation--what are we measuring? *Med Teach*. 2009;31(2):e36-e39. doi:10.1080/01421590802334309
6. Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Med Educ*. 2016;16(1):266. doi:10.1186/s12909-016-0793-z
7. Arora C, Sinha B, Malhotra A, Ranjan P. Development and Validation of Health Education Tools and Evaluation Questionnaires for Improving Patient Care in Lifestyle Related Diseases. *J Clin Diagn Res*. 2017;11(5):JE06-JE09. doi:10.7860/JCDR/2017/28197.9946
8. Badyal DK, Singh S, Singh T. Construct validity and predictive utility of internal assessment in undergraduate medical education. *Natl Med J India*. 2017;30(3):151-154.
9. Marceau M, Gallagher F, Young M, St-Onge C. Validity as a social imperative for assessment in health professions education: a concept analysis. *Med Educ*. 2018;52(6):641-653. doi:10.1111/medu.13574
10. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. *Adv Med Educ Pract*. 2017;8:567-570. doi:10.2147/AMEP.S139492
11. Sennekamp M, Gilbert K, Gerlach FM, Guethlin C. Development and validation of the "FrOCK": frankfurt observer communication checklist. *Z Evid Fortbild Qual Gesundhwes*. 2012;106(8):595-601. doi:10.1016/j.zefq.2012.07.018
12. Young M, St-Onge C, Xiao J, Vachon Lachiver E, Torabi N. Characterizing the literature on validity and assessment in medical education: a bibliometric study. *Perspect Med Educ*. 2018;7(3):182-191. doi:10.1007/S40037-018-0433-X
13. Abma IL, Rovers M, van der Wees PJ. Appraising convergent validity of patient-reported outcome measures in systematic reviews: constructing hypotheses and interpreting outcomes. *BMC Res Notes*. 2016;9(1):226. doi:10.1186/s13104-016-2034-2