


The Reliability of 2-Station Clerkship Objective Structured Clinical Examinations in Isolation and in Aggregate

Journal of Medical Education and
Curricular Development
Volume 6: 1–7
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2382120519863443



Aaron W Bernard¹ , Richard Feinn², Gabbriel Ceccolini³,
Robert Brown⁴, Ilene Rosenberg², Walter Trymbulak⁵
and Christine VanCott⁶

¹Department of Medical Sciences, Frank H. Netter MD School of Medicine, Quinnipiac University, Hamden, CT, USA. ²Department of Medical Sciences, Frank H. Netter MD School of Medicine, Quinnipiac University, Hamden, CT, USA. ³Standardized Patient and Assessment Center, Frank H. Netter MD School of Medicine, Quinnipiac University, Hamden, CT, USA. ⁴Department of Medicine, Frank H. Netter MD School of Medicine, Quinnipiac University, Hamden, CT, USA. ⁵Department of Obstetrics and Gynecology, Frank H. Netter MD School of Medicine, Quinnipiac University, Hamden, CT, USA. ⁶Department of Surgery, Frank H. Netter MD School of Medicine, Quinnipiac University, Hamden, CT, USA.

ABSTRACT

BACKGROUND: Most medical schools in the United States report having a 5- to 10-station objective structured clinical examination (OSCE) at the end of the core clerkship phase of the curriculum to assess clinical skills. We set out to investigate an alternative OSCE structure in which each clerkship has a 2-station OSCE. This study looked to determine the reliability of clerkship OSCEs in isolation to inform composite clerkship grading, as well as the reliability in aggregate, as a potential alternative to an end-of-third-year examination.

DESIGN: Clerkship OSCE data from the 2017-2018 academic year were analyzed: the generalizability coefficient (ρ^2) and index of dependability (ϕ) were calculated for clerkships in isolation and in aggregate using variance components analysis.

RESULTS: In all, 93 students completed all examinations. The average generalizability coefficient for the individual clerkships was .47. Most often, the largest variance component was the interaction between the student and the station, indicating inconsistency in the performance of students between the 2 stations. Aggregate clerkship OSCE analysis demonstrated good reliability for consistency ($\rho^2 = .80$). About one-third (33.8%) of the variance can be attributed to students, 8.2% can be attributed to the student by clerkship interaction, and 42.6% can be attributed to the student by block interaction, indicating that students' relative performances varied by block.

CONCLUSIONS: Two-station clerkship OSCEs have poor to fair reliability, and this should inform the weighting of the composite clerkship grade. Aggregating data results in good reliability. The largest source of variance in the aggregate was student by block, suggesting testing over several blocks may have advantages compared with a single day examination.

KEYWORDS: OSCE, clerkship, reliability, assessment, medical students, USMLE, step 2 CS

RECEIVED: January 31, 2019. **ACCEPTED:** June 24, 2019.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Aaron W Bernard, Department of Medical Sciences Frank H. Netter MD School of Medicine, Quinnipiac University, 275 Mt. Carmel Avenue, Hamden, CT 06518, USA.
Email: aaron.bernard@quinnipiac.edu

Introduction

The objective structured clinical examination (OSCE) is a commonly used assessment method to test students' clinical skills in medical schools in the United States. The examinations can take on several forms, but commonly consist of a series of stations in which a student performs a history and physical on a standardized patient (SP) and then writes a medical note. Through this process, a student is assessed on a variety of skills including history taking, physical examination, communication, medical documentation, and clinical reasoning.

The psychometric properties of this form of assessment are complex and nuanced. It is clear that the reliability of the examination increases with the number of stations.^{1–3} The

United States Medical Licensing Examination Step 2 Clinical Skills (USMLE Step 2 CS) uses 12 stations, a few of which are unscored pilot cases, to reliably discriminate about pass/fail cut scores for 3 domains.^{3,4} Most US medical schools report having a high-stakes, end-of-third-year OSCE which is often linked to advancement, graduation, and/or permission to sit for the USMLE Step 2 CS examination.⁵ These examinations frequently consist of 5 to 10 stations.^{2,6,7} Objective structured clinical examinations of this length can be impractical to administer with frequency throughout a medical school curriculum because of the time needed and resources required.

In 2015, the Frank H. Netter MD School of Medicine (Netter SOM) had its first cohort of students enter the



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<http://www.creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Table 1. Structure of the composite clerkship grade.

CLERKSHIP ASSESSMENT COMPONENT	PERCENTAGE OF COMPOSITE CLERKSHIP GRADE
Clinical evaluations by faculty	40
NBME subject examination	40
OSCE	10
Clerkship specific assessment (modules, case write-ups, case presentation, etc.)	10

Abbreviations: NBME, National Board of Medical Examiners; OSCE, objective structured clinical examination.

third-year clerkship phase of the curriculum. In designing that phase, the school faculty decided to have an OSCE for each clerkship to help ensure a robust composite clerkship grade.^{8,9} However, it was operationally prohibitive to have 5- to 10-station OSCEs for all 6 clerkships, each of which repeat 7 times per year. A 2-station clerkship OSCE structure was deemed feasible and adopted with the OSCE contributing 10% of the composite clerkship grade. This percentage was chosen by faculty in a consensus manner not only attempting to ensure that the examination weight had significance but also taking into account the expected reliability of a short OSCE. As an added benefit of this clerkship OSCE structure, the school intended the frequent OSCEs to provide longitudinal practice for the USMLE Step 2 CS examination.

There are no prior reports in the literature of a US medical school that runs an OSCE for every third-year clerkship. As such, this study set out to examine several outcomes regarding our innovative curricular design by examining a cohort of 93 students. The first was to determine the actual reliability of the 2-station OSCEs to inform our composite clerkship assessment structure and provide insights for other medical schools. The second was to look at third-year clerkship OSCE data in aggregate. This study aimed to determine reliability of such aggregate data so as to potentially serve in the future as a substitute for a more traditional, long, end-of-third-year OSCE.¹⁰ Would we be able to reliably assess student performance in this manner with the intent to identify students who need extra coaching prior to graduation and/or the USMLE Step 2 CS examination?¹⁰ Finally, we aimed to publicly report our school's relevant USMLE performance data for other schools considering a similar third-year clerkship OSCE structure to review.

Methods

Study setting and participants

The Quinnipiac University institutional review board determined this study to be exempt from review. This study took place at the Netter SOM, which is a community-based medical school with an enrollment of 90 to 96 students per class. The third-year clerkship phase consists of 6 clerkships (internal medicine, obstetrics and gynecology, pediatrics, primary care, psychiatry, and surgery) and one required block in which the

students can choose from a few non-clinical educational experiences such as research. Students take these clerkships in varying order, as is the norm at most medical schools, to distribute students and maximize clinical teaching capacity. All clerkships are 6 weeks in length. The structure of the clerkship composite grade is described in Table 1.

Every clerkship has a 2-station OSCE completed in the last week of the clerkship that is formatted to mimic the structure of the USMLE Step 2 CS examination. Students are given 15 minutes to perform a history and physical, and then 10 minutes to complete a note that involves documentation of their history, physical, differential with supporting evidence, and recommended testing.⁴ Although the structure of the OSCE remains stable across all clerkships, the case content is aligned with the learning objectives of the associated clerkship. The OSCE scoring is described in Table 2.

Objective structured clinical examination case design follows an iterative process involving the clerkship director, clerkship faculty, and the medical director of the clinical skills center. Case content ideas are derived from clerkship learning objectives, and this is followed by case and checklist development. Cases are next made live in the clerkship OSCE and then undergo continuous quality improvement involving SP, faculty, and student feedback. In addition, case difficulty is periodically reviewed and checklists are reviewed using case-item analysis. The first year for the cases was 2015-2016. This study represents the 2017-2018 academic year—a time by which all cases had been live and undergone continuous quality improvement for at least 1 year.

The Netter SOM uses formative OSCEs with 20-minute stations in the first 2 years of the curriculum.¹¹ Students get immediate feedback from SPs after each station, are allowed to review their videos and checklists, and have dedicated video review time with faculty. Those OSCEs have a similar history, physical, and communication checklist structure. Students and SPs are therefore familiar with this assessment structure at the start of the clerkship OSCEs. Experienced SPs are selected for clerkship OSCEs given the summative nature, and whenever possible, the same SPs are used for all blocks of a specific clerkship. All SPs have case portrayal and scoring accuracy periodically reviewed and are provided with feedback by an SP educator.

Clerkship directors, serving as content experts, anonymously score their clerkships' notes each block. Scoring follows a rubric

Table 2. Structure of the OSCE score.

Total OSCE score = $\frac{\text{Case 1 score} + \text{Case 2 score}}{2}$	
Case score = $\frac{A + B + C + D}{4}$	
A	SP scored History Checklist=yes/no Thoroughness items such as “asked about quality of pain” or “asked about past medical history.”
B	SP scored Physical Examination Checklist=correct/incorrect/not done
C	SP scored Master Interview Rating Scale=Scaled 1-5 Communication items such as introduction, types of questions, non-verbal behavior, empathy, achieve a shared plan.
D	Faculty scored Post-Encounter Note

Abbreviations: OSCE, objective structured clinical examination; SP, standardized patient.

created by a consensus process with consideration given to published rubrics.¹² See Table 3 for the Netter clerkship OSCE note scoring rubric. Students are provided a report of their overall score for each of the clerkship OSCEs as well as a score by skill assessed. In contrast to the pre-clerkship formative OSCEs, students do not get immediate SP feedback and are not provided access to videos or checklists on clerkship OSCEs for examination security reasons. During the study period, students were permitted to ask for additional feedback from the clerkship director or the medical director of the clinical skills center.

Data abstraction

The simulation management software used to run OSCEs at the Netter SOM was accessed and all third-year clerkship OSCE data from the 2017-2018 academic year were confidentially abstracted for analysis. The National Board of Medical Examiners (NBME) does not provide individual student reports for the USMLE Step 2 CS examination to medical schools but instead provides an annual school wide performance report. All reports since the inception of the Netter SOM (2015-2016, 2016-2017, 2017-2018) were confidentially obtained from the Netter SOM Office of Assessment.

Data analysis

To compare OSCE total scores between blocks and between clerkships, a 2-factor linear mixed model with fixed effects for block and clerkship and a random intercept for student were used with post hoc Bonferroni paired comparisons if a significant effect was found. A variance components analysis was used to estimate the reliability of the OSCE score by clerkship. Separately for each clerkship, a model was run with student and station as random effects. It was decided to not run models that included additional sources of variation such as SP and block because of model complexity. Using generalizability theory, the generalizability coefficient (ρ^2) measuring consistency and index of dependability (ϕ) measuring absolute agreement

were calculated. Variance components were then used to assess how much of the variability in OSCE scores across the entire third year can be attributed to variation in student performance, variation between blocks, variation between clerkships, and the interaction effects among these factors. For this analysis, the average of the 2 stations constituted the OSCE score. Employing generalizability theory, the calculated variance components were then used to estimate the reliability of taking the average OSCE scores across all clerkships. Students were crossed with blocks and crossed with clerkships, and blocks were crossed with clerkships; however, this was not fully crossed in that there was not a student by block by clerkship (3-way interaction) effect. All analyses were conducted in SPSS v24 and the alpha level for statistical significance was set at .05.

The USMLE Step 2 CS data are reported as groups of students by academic year (June-July) which does not perfectly correspond to the time the cohort of students of this study took the examination (April-December). As such, the authors' consensus decision was to summarize the general trends of the 3 available reports.

Results

Ninety-three students completed the third-year curriculum in the 2017-2018 academic year. All students had OSCE data for all 6 clerkships. Table 4 shows the average percentage OSCE total score by block and clerkship. There was no significant interaction between block and clerkship ($P = .10$), and the average scores were consistent across the 7 blocks and showed no significant difference ($P = .13$). There was a significant difference in scores between the 6 clerkships ($P < .001$) with obstetrics and gynecology having the lowest average and psychiatry the highest. Post hoc Bonferroni-paired comparisons revealed OSCE scores for psychiatry were significantly higher than all other clerkships, obstetrics and gynecology were significantly lower than all other clerkships except internal medicine, and internal medicine scores were significantly lower than primary care. To better understand the source of the significant difference between clerkships, further analyses

Table 3. Netter clerkship OSCE note scoring rubric.

SECTION	POINTS	RUBRIC
History	4	4: All key information present, well organized.
		3: A single key information item missing, somewhat disorganized.
		2: Several key information items missing, disorganized.
		1: Key information is incorrect, very disorganized.
		0: Nothing recorded.
Physical examination	4 N/A for pediatrics	4: All key information present, well organized. ^{a,b}
		3: A single key information item missing, somewhat disorganized.
		2: Several key information items missing, disorganized.
		1: Key information is incorrect, very disorganized.
		0: Nothing recorded.
Differential and evidence	4	4: All diagnoses listed in correct order with correct and complete supporting evidence.
		3: All diagnoses listed but not in correct order. A few missing or incorrect supporting evidence elements.
		2: One expected diagnoses missing. Several missing or incorrect supporting evidence elements.
		1: All expected diagnoses missing.
		0: Nothing recorded.
Testing	4	4: Plan for diagnostic workup is effective and efficient, includes all essential tests, and few or none unnecessary tests. ^c
		3: Reasonable plan for diagnostic workup, may have some unnecessary tests or missing few essential tests. Lists therapy or consultations as part of diagnostic workup.
		2: Ineffective plan for diagnostic workup—essential tests missed, irrelevant tests included.
		1: Diagnostic workup places patient in unnecessary risk or danger. (unjustified invasive procedures)
		0: Nothing recorded.

Abbreviation: OSCE, objective structured clinical examination.

^aVital signs are considered key information.

^bMental status examination is included as key information in the physical examination section on the Psychiatry OSCE, in addition to the vitals and traditional physical examination.

^c"No diagnostic testing indicated" is appropriate in some cases and can result in a score of 4/4. This is different than leaving the section blank.

compared the clerkships on the 4 subscales that compose the total OSCE score (Table 5). There was a significant difference between clerkships on history ($P < .001$), communication ($P < .001$), and note taking ($P < .001$), but not physical examination ($P = .57$).

Table 6 shows the variance components along with the reliability coefficients for each clerkship. The reliabilities ranged from poor (pediatrics) to fair (psychiatry) with the absolute agreement coefficients (ϕ) being just slightly lower than the consistency coefficients (ρ^2). The average generalizability coefficient across the 6 clerkships was .47 and the average index of dependability was .45. Most often, the largest variance component was the interaction between the student and the station, indicating inconsistency in the relative performance of students

between the 2 stations, and the smallest variance component was the station, indicating little variance due to varying levels of difficulty between stations.

Table 7 shows the variance components analysis for the aggregate OSCE total score across the entire year 3 curriculum. About one-third (33.8%) of the variance in scores can be attributed to the students. The largest source of variation is from the student by block interaction (42.6%), indicating students' relative performances varied by block. The student by clerkship interaction (8.2%) also showed relative performance varied somewhat by clerkship. Taking the aggregate of the 6 clerkship OSCE scores as a measure of student performance results in good reliability for both consistency ($\rho^2 = .80$) and absolute agreement ($\phi = .75$).

To date, including most of the students from the 2017-2018 academic year analyzed in this study, the Netter SOM has a 100% pass rate on the USMLE Step 2 CS examination (N = 223). The NBME provides medical schools with graphical reports of performance in the various subcategories of the examination. A rough visual analysis of 3 years of reports from the NBME suggests the Netter SOM typically performs around 2 standard deviations above the mean for communication and between 0.5 and 1.0 standard deviation above the mean for the integrated clinical encounter component of the examination.⁴

Table 4. Average percentage OSCE score by block and clerkship (N=93 students).

SOURCE	MEAN	SD
Block		
1	88.9	5.7
2	89.8	5.8
3	88.8	5.7
4	89.1	5.2
5	90.3	4.2
6	88.8	4.9
7	89.4	5.2
Clerkship		
Internal medicine	88.1	4.9
OB/GYN	86.6	5.1
Primary care	90.4	4.6
Pediatrics	88.9	4.9
Psychiatry	92.6	4.6
Surgery	89.2	5.4

Abbreviations: OB/GYN, obstetrics and gynecology; OSCE, objective structured clinical examination.

Discussion

There are no prior reports in the literature of a US medical school that runs an OSCE for every third-year clerkship in a manner such as the Netter SOM—ensuring consistent OSCE structure, length, and assessment instruments. This is in part related to the administrative burden required to run OSCEs for every clerkship. Traditional, 5- to 10-station OSCEs are time-consuming and expensive.^{2,6,7} This study describes the reliability of a more feasible 2-station clerkship OSCE at 1 institution for 6 clerkships. The data are interesting in that the reliability for some clerkships (internal medicine, pediatrics, surgery) appears poor while the reliability for other (obstetrics and gynecology, primary care, psychiatry) appears fair. Student performance was the second largest source of variance. Most of the variance came from student by station interaction—some students did better on station A while others did better on station B.

A common explanation for student by station variance is that students master different patient types during their clerkship—either because of differences in clinical exposure or differences in what students study.² However, this theory does not fully explain why some clerkships seem to be impacted by this phenomenon more than others. We hypothesize that at our institution, the length of the clerkships may have contributed to the finding. All of the clerkships at Netter SOM are 6 weeks. It may be that for some clerkships like internal medicine and surgery, which are longer at most institutions, there is simply not enough time for students to be exposed to and/or study all the patient types.¹³ This explanation also makes sense for pediatrics in which OSCE stations present children of various ages. Another possible explanation for the varied student performance on pediatric cases could be the atypical interview students must perform. Similar to the USMLE Step 2 CS examination, student do not actually evaluate children—they either interview a mother of a child in person or they are talking to a parent on the phone about their child.⁴

The average reliability of the clerkship 2-station OSCEs was somewhat low (.47). However, we hope our work helps to reframe the discussion around the reliability of OSCEs. We

Table 5. Average percentage (SD) OSCE subscale score by clerkship (N=93 students).

SOURCE	HISTORY	PHYSICAL EXAMINATION	COMMUNICATION	NOTES
Clerkship				
Internal medicine	95.8 (4.4)	86.3 (12.6)	90.9 (7.3)	79.4 (7.7)
OB/GYN	97.2 (5.5)	86.7 (10.7)	89.2 (7.5)	73.3 (9.6)
Primary care	91.0 (7.3)	87.4 (10.8)	90.9 (7.2)	92.3 (3.1)
Pediatrics	88.8 (7.7)		87.7 (8.4)	90.2 (8.7)
Psychiatry	93.2 (7.2)	86.9 (12.5)	93.2 (6.3)	97.0 (3.5)
Surgery	89.2 (5.4)	88.6(11.1)	92.7 (6.6)	80.9 (10.4)

Abbreviations: OB/GYN, obstetrics and gynecology; OSCE, objective structured clinical examination.

Table 6. Variance components (%) for OSCE total score by clerkship.

SOURCE	IM	OB/GYN	PC	PEDS	PSYCH	SURG
Student	7.1 (17%)	17.9 (43%)	11.6 (33%)	5.5 (13%)	14.8 (54%)	10.7 (22%)
Station	2.5 (6%)	4.9 (12%)	2.4 (7%)	2.4 (5%)	1.1 (4%)	3.7 (8%)
Student × Station	32.7 (77%)	18.5 (45%)	20.7 (60%)	36.5 (83%)	11.7 (43%)	33.6 (70%)
Index of dependability (φ)						
Average across 2 stations	.288	.605	.501	.222	.698	.364
Generalizability coefficient ($E\rho^2$)						
Average across 2 stations	.304	.660	.528	.232	.716	.388

Abbreviations: IM, internal medicine; OB/GYN, obstetrics and gynecology; OSCE, objective structured clinical examination; PC, primary care; PEDS, pediatrics; PSYCH, psychiatry; SURG, surgery.

Table 7. Variance components for aggregate OSCE total score.

SOURCE	VARIANCE	% VARIANCE
Student	9.56	33.8
Block	0.05	0.2
Clerkship	3.90	13.8
Student × Block	12.02	42.6
Student × Clerkship	2.32	8.2
Block × Clerkship	0.44	1.6
Index of dependability (φ)		
Average across clerkships	.754	
Generalizability Coefficient ($E\rho^2$)		
Average across clerkships	.800	

Abbreviation: OSCE, objective structured clinical examination.

recommend clerkship faculty not to look at reliability calculations in isolation but rather consider the sufficiency of the reliability when placed in the context of the contribution to the composite clerkship grade.⁷ Our faculty are content with the 2-station OSCE accounting for 10% of the current composite clerkship grade. The importance of varied assessment components in a composite clerkship grade is reinforced by our research as well as recent reports suggesting faculty clinical evaluations lack consistency of reliability across clerkships.¹⁴ Excellent reliability with all of the common assessment methods used in clerkships may be difficult to achieve.¹⁵ The NBME subject examinations have better reliability, but if used in isolation for a clerkship grade, validity would be missing as medical knowledge is not the only competency that must be evaluated.⁸

The Netter SOM stands in the minority of US medical school curricula as not having an end-of-third-year OSCE.⁵ One of the many arguments in favor of such an examination is to ensure students have accumulated desired clinical skills prior

to advancement to the next stage of the curriculum. Netter SOM faculty felt this could be done within the context of clerkship assessments and did not appreciate enough added value in an additional stand-alone assessment point that justified the time, resources, and effort. Netter SOM faculty did, however, recognize that an end-of-third-year OSCE could be of value, not in preventing inappropriate advancement but rather in identifying students who could benefit from additional coaching in their final year of training. It has always been our hope to pull our third-year OSCE data out of the individual clerkships, aggregate it, and use it for such purpose. This study was our first attempt to aggregate data, and we did establish a high degree of reliability in doing so ($\rho^2 = .80$).

Moving forward, our protocol is going to be to notify the bottom 10% of performers of the aggregate data that coaching is advised and assign a clinical skills coach to work with the student prior to sitting for the USMLE Step 2 CS examination. The advantage of this approach is we are targeting limited resources to students who are most in need of help. The disadvantage of this approach is that it may be overly focused on the examination. All students may benefit from coaching for clinical skill development, even those not in danger of failing the Step 2 CS examination. In addition, we are re-considering our feedback process throughout the third year. Initially, we were concerned about examination security and therefore only provided limited feedback to students. Additional OSCE cases are under development for the clerkships. With a larger case bank that is periodically rotated, examination security may be less of an issue, and we may provide more robust feedback similar to our pre-clerkship formative OSCEs.

The sources of variance in the aggregate data were mostly student and student by block but not so much student by clerkship. This suggests that while some students perform better than others do, individual students have stronger and weaker performances throughout the year. These data make the argument that it may be ideal to gather data longitudinally and then aggregate the data as was done in this study. A one-time end-of-third-year OSCE has the advantage of capturing the

students' performance at one particular point in time, but our works raise the possibility that a student's performance may be better measured through a longitudinal assessment capturing several data points over time.

As noted above, the Netter curriculum lacks a comprehensive end-of-third-year examination and stands in the minority of US medical schools in this regard.⁵ As such, we did want to report school-wide performance data from the USMLE Step 2 CS examination. Over several years now, we have maintained a 100% pass rate. In general, our school outperforms the national average in both the communication and integrated clinical encounter sections. This study was not designed to determine the reasons for that success but we have a hypothesis. The literature on re-take USMLE Step 2 CS examinations suggests that part of performance is familiarity with the structure of the examination.¹⁶ The Netter SOM OSCE structure allows for repetitive exposure and practice within to the structure of the USMLE examination format.

Limitations

This is a single-center study, and the reliability may have institutional factors that limit generalizability to other medical schools. We recommend routine analysis of local medical school OSCE data. The study also represents the data from 1 cohort of students. Although we do have data from 2015–2016 and 2016–2017, we do not have as much confidence in that data and felt more comfortable analyzing data from the 2017–2018 academic year which had the luxury of 2 previous years of continuous quality improvement.

The Step 2 CS data are only one measure of curriculum effectiveness. It would be valuable to gather student feedback in both qualitative and quantitative ways. Do students feel comfortable taking the 10-station Step 2 CS examination, having primarily practiced with 2-station OSCEs? Do students feel prepared for the Step 2 CS examination? Would students recommend any changes in the curriculum to aid in preparation? Putting the Step 2 CS examination aside, do students feel the clerkship OSCEs would have been more valuable to their clinical skill development if detailed feedback was provided? Should the clerkship curriculum include formative OSCEs to ensure that assessment for learning is occurring?

Conclusions

Two-station clerkship OSCEs have poor-to-fair reliability and this should inform the weighting of the composite clerkship grade. Aggregating data from 6 clerkship OSCEs results in good reliability. The largest sources of variance in the aggregate were student and student by block, suggesting testing over several blocks may have advantages over a single day examination.

Acknowledgements

We would like to recognize Jean Kappes, Lauren Kowalewski, and Joel Morgenstern for their valued contributions to this investigation.

Author Contributions

All authors made a substantial contribution to the design of the work, drafted or critically revised the manuscript, approved the final version, and took public responsibility for the work.

Ethical Approval

The Quinnipiac University institutional review board determined this study to be exempt from review.

ORCID iD

Aaron W Bernard  <https://orcid.org/0000-0002-4065-6579>

REFERENCES

1. Brannick MT, Erol-Kormaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ.* 2011;45:1181–1189.
2. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med.* 2008;40:574–578.
3. Burdick WP, Boulet JR, LeBlanc KE. Can we increase the value and decrease the cost of clinical skills assessment? *Acad Med.* 2018;93:690–692.
4. United States Medical Licensing Examination. Step 2 clinical skills (CS) (Content description and general information). <https://www.usmle.org/pdfs/step-2-cs/cs-info-manual.pdf>. Accessed September 1, 2018.
5. Association of American Medical Colleges. Number of medical schools requiring final SP/OSCE examination: 2011–2012 through 2015–2016. <https://www.aamc.org/initiatives/cir/406426/9.html>. Accessed September 1, 2018.
6. Park YS, Lineberry M, Hyderi A, Bordage G, Xing K, Yudkowsky R. Differential weighting for subcomponent measures of integrated clinical encounter scores based on the USMLE Step 2 CS examination: effects on composite score reliability and pass-fail decisions. *Acad Med.* 2016;91:S24–S30.
7. Berg K, Winward M, Clauser BE, et al. The relationship between performance on a medical school's clinical skills assessment and USMLE Step 2 CS. *Acad Med.* 2008;83:S37–S40.
8. Corcoran J, Downing SM, Tekian A, DaRosa D. Composite score validity in clerkship grading. *Acad Med.* 2009;84:S120–S123.
9. Schilling DC. Using the clerkship shelf exam score as a qualification for overall clerkship grade of honors: a valid practice or unfair to students? *Acad Med.* 2019;94:328–332.
10. Bergus G, Kreiter CD. The reliability of summative judgements based on objective structured clinical examination cases distributed across the clinical year. *Med Educ.* 2007;41:661–666.
11. Bernard AW, Ceccolini G, Feinn R, et al. Medical students review of formative OSCE scores, checklists, and videos improves with student-faculty debriefing meetings. *Med Educ Online.* 2017;22:1324718.
12. Yudkowsky R, Park YS, Hyderi A, Bordage G. Characteristics and implications of diagnostic justification scores based on the new patient note format of the USMLE step 2 CS exam. *Acad Med.* 2015;90:S56–S62.
13. Association of American Medical Colleges. Required weeks for: surgery. <https://www.aamc.org/initiatives/cir/426756/05e.html>. Accessed August 19, 2018.
14. Zaidi NLB, Kreiter CD, Castaneda PR, et al. Generalizability of competency assessment scores across and within clerkships: how students, assessors, and clerkships matter. *Acad Med.* 2018;93:1212–1217.
15. Hauer K, Lucey CR. Core clerkship grading: the illusion of objectivity. *Acad Med.* 2019;94:469–472.
16. Swygert KA, Balog KP, Jobe A. The impact of repeat information on examinee performance on a large-scale standardized-patient examination. *Acad Med.* 2010;85:1506–1510.