The American Journal of Surgery (2014) 207, 236-242

The American Journal of Surgery®

Association for Surgical Education

Validity and reliability of a novel written examination to assess knowledge and clinical decision making skills of medical students on the surgery clerkship

Anna Reinert, B.A.[a], Ana Berlin, M.D.[a], Aubrie Swan-Sein, Ph.D., M.Ed.[b], Roman Nowygrod, M.D.[a], Abbey Fingeret, M.D.[a,*]

[a]Department of Surgery, New York-Presbyterian Hospital, Columbia University Medical Center, 177 Fort Washington Avenue, MHB 7GS-313, New York, NY 10032, USA; [b]Center for Education Research and Evaluation, Columbia University Medical Center, New York, NY, USA

KEYWORDS:
Surgical education;
Medical student;
Clinical skills;
Clinical reasoning;
Objective examination;
Assessment

Abstract

BACKGROUND: The Surgery Clerkship Clinical Skills Examination (CSE) is a novel written examination developed to assess the surgical knowledge, clinical decision making, communication skills, and professionalism of medical students on the surgery clerkship. This study was undertaken to determine its validity.

METHODS: Data were prospectively collected from July 2011 through February 2013. Multivariate linear and logistic regression analyses were used to assess score trend; convergent validity with National Board of Medical Examiners surgery and medicine subject scores, United States Medical Licensing Examination Step 1 and Step 2 Clinical Knowledge scores, and evaluation of clinical reasoning and fund of knowledge; and the effect of clerkship order. Exam reliability was assessed using a modified Cronbach's α statistic.

RESULTS: During the study period, 262 students completed the CSE, with a normal distribution of performance. United States Medical Licensing Examination Step 2 Clinical Knowledge score and end-of-clerkship evaluations of fund of knowledge and clinical reasoning predicted CSE score. Performance on the CSE was independent of clerkship order or prior clerkships. The modified Cronbach's α value for the exam was .67.

CONCLUSIONS: The CSE is an objective, valid, reliable instrument for assessing students on the surgery clerkship, independent of clerkship order.

Objective appraisal of student performance on the surgery clerkship has traditionally relied on the National Board of Medical Examiners (NBME) clinical surgery subject examination or "shelf." This exam, consisting of multiple-choice questions, is a valuable instrument in student assessment but one that measures primarily a single dimension of student performance: achievement in the domain of surgical knowledge. Medical knowledge is 1 of 6 Accreditation Council for Graduate Medical Education core competencies put forth as outcome-based standards in medical education.[1–3] Although

* Corresponding author. Tel.: +1-212-305-5970; fax: +1-212-305-8321.
E-mail address: af2451@columbia.edu

Manuscript received May 6, 2013; revised manuscript August 20, 2013

0002-9610/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amjsurg.2013.08.024

medical knowledge can be easily, effectively, and reliably assessed through a standardized format of multiple-choice questions, the other 5 core competencies describe behaviors and habits that require alternative methods of evaluation.[2,4,5]

Medical education theory classifies methods of trainee assessment into 4 categories of achievement: "knows," "knows how," "shows," and "does," depicted by Miller's pyramid (Fig. 1).[6] Multiple-choice questions assess what a medical trainee knows, or the trainee's ability to recognize the correct answer from a list of possible responses. The cognitive skill measured in this way is different from clinical decision making, which is a skill better assessed through open-response, task-based exam formats corresponding to the "knows how" level of Miller's pyramid. Written case-based simulation is one such method of "knows how" assessment.[6,7] This exam format allows the sampling of a large number of clinical topics within a single exam administration, testing essential elements in decision making and critical steps in the successful resolution of the clinical problem; clinical judgment or reasoning and problem solving abilities of examinees are measured with professional realism.[8–10] Compared with multiple-choice question exams, these exams are less influenced by cueing and do not overestimate examinees' ability.[11] The major limitation of these examinations is case specificity, which results in lower reliability than may be achieved with multiple-choice question exams.[5,7]

Survey data from 2007 and 2008 show that the NBME clinical surgery subject examination is used by 90% to 95% of medical school surgery programs in the United States and is given an average weight of 31% in the determination of final clerkship grade.[12,13] Of surgery programs surveyed by the NBME, 99% report that they are somewhat to very satisfied with the ability of the subject exam to evaluate students' knowledge. Yet only 29% of surgery programs surveyed reported themselves to be more than somewhat satisfied with school curriculum match to subject exam content, and 4% reported being not at all satisfied.[12] As the NBME's *Surgery Subject Examination Score Interpretation Guide* states, "Subject examination scores should not be used alone, but rather in conjunction with other indicators of examinee performance in determination of grade."[14] Of the medical school surgery programs surveyed, 37.7% reported the use of an objective structured clinical examination, and 37.7% reported the use of a written exam other than the NBME clinical surgery subject examination.[13]

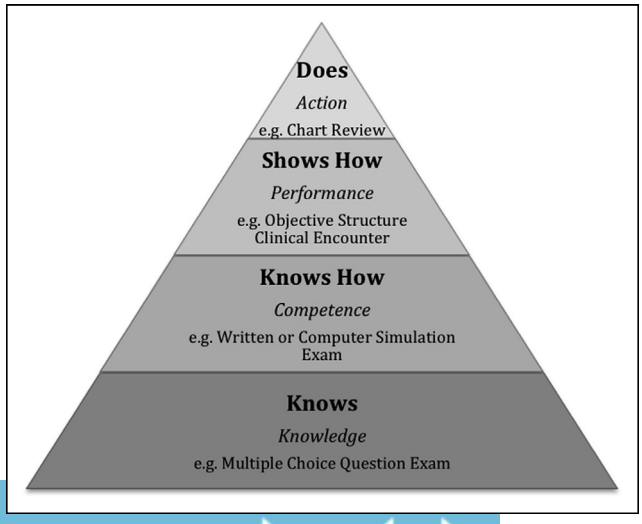

Figure 1 Miller's pyramid framework for clinical assessment. Adapted with permission from Miller.[6]

Locally developed written exams have the benefits that they can be tailored to medical school educational objectives and can be designed to differentiate within a peer group.[15] Toward this end, the Surgery Clerkship Clinical Skills Examination (CSE) was developed at our institution. The goals of exam development were to create an objective, reliable, validated instrument for assessing students' skill in applying surgical knowledge to clinical scenarios and to assess the additional Accreditation Council for Graduate Medical Education competencies of patient care, communication, and professionalism, educational objectives that are not appropriately measured by the NBME clinical surgery subject examination. The CSE also served to replace unstructured oral exams that were traditionally used on the surgery clerkship at our institution but were discontinued because of resource-intensiveness, subjectivity in grading, and student feedback suggesting a negative experience with the evaluation process.

An additional issue with the NBME surgery subject exam is that scores have been observed to trend seasonally, with students who complete the surgery clerkship later in the clinical year performing significantly better than those who complete the surgery clerkship earlier in the course of the clinical year.[14,16–18] The presence of seasonal variation in scores results in a lack of comparability between students completing the surgery clerkship at different points in the clinical year, raising questions about fairness in grade assignment.[16] We hypothesized that scores on the CSE would not exhibit such a seasonal trend, because of general surgery case specificity with less overlap with the subject areas of medicine, surgical subspecialties, and obstetrics and gynecology.

Methods

Exam development

The CSE is a written, case-based, clinical simulation exam composed of 5 case scenarios per exam drawn from a pool of >20 scenarios. A dedicated surgical education research fellow developed the scenarios with faculty consensus regarding content and scoring rubric. Each scenario constitutes 100 points and is scored using a detailed keyword rubric by senior residents after training and establishment of interrater reliability >.90. Graders are trained with a formal orientation followed by practice

grading of selected student responses. After this training period, remediation is provided if initial reliability is <.90 on sample items. Volunteer resident graders, including 3 members of each of the 4th and 5th postgraduate year classes, are compensated $40 per hour for training and grading. Each student examination takes roughly 1 hour to grade. After the pilot period, including costs for question development, resident training and grading, exam posting, score collating and review, and quarterly quality assurance with algorithm-based statistical analysis, the overall examination administration expense is $55 per student. The NBME subject examination fee is $40 per student.[19]

The CSE format consists entirely of free-response questions: short-answer, long-answer essay, and algorithm completion. Scenarios are linear rather than branching: content does not change on the basis of students' responses to earlier questions. Rather, additional information is provided about the case as the student proceeds to successive sections, and responses to earlier questions cannot be revised. This format allows students to revise incorrect thinking about the case in their responses to later questions, such that errors in first-order clinical reasoning do not determine performance on higher order questions. Notable is the realism of this exam, which asks students to think clinically and make decisions mirroring those commonly required of practicing surgeons.

Each scenario variably assesses knowledge of key features of the history and physical exam findings pertinent to a clinical presentation, formulation of a differential diagnosis, interpretation of radiographic and photographic evidence, knowledge of surgical anatomy, and recognition and management of common postoperative complications. The exam also assesses communication skills by simulating the description of clinical course, expectations, informed consent, or difficult news to the mock patient. Professionalism is evaluated with content addressing ethical principles pertaining to the provision or withholding of clinical care. The exam is administered over 3 hours by computer using SofTest Software (ExamSoft Worldwide, Inc, Dallas, TX).

After the examination, students are invited to attend an optional session reviewing the exam content and grading rubric and asked to provide informal feedback on the content relative to course objectives, clarity, relevance, and perceived fairness. Additionally, students may provide anonymous evaluations of the clerkship, including four 9-point Likert scale items related to the exam on a spectrum from "does not meet" to "exceeds expectations": relation to course objectives, relevance and realism of clinical scenarios, clarity of questions, and perceived opportunity to provide feedback. Students may also provide free-text comments on any aspect of the exam. This informal and formal feedback is used in composite with specific scenario aggregate performance to revise or modify test items. The CSE constitutes 30% of the overall course evaluation, with the NBME surgery subject examination accounting for 10% and the remaining 60% from a composite clinical evaluation. The exam is curved by individual block, with a minimum requirement of performance above the mean to obtain a grade of honors.

Data collection

Students on the surgery clerkship take the CSE and NBME surgery subject examination as required components of their end-of-clerkship assessments. The Department of Surgery maintains individual student data on CSE total score and subtotal scores for each scenario. The institutional Center for Educational Research & Evaluation maintains a database of individual student performance throughout medical school, which includes the parameters of United States Medical Licensing Examination (USMLE) Step exam scores, NBME subject exam scores, and resident, faculty, and clerkship director evaluative ratings of students' clerkship performance as collected using software (New Innovations, Uniontown, OH).

After a pilot period, and after institutional review board study approval, prospective data collection was carried out for students taking the CSE during the surgery clerkship from July 2011 through February 2013. Data for each student were matched to the Center for Educational Research & Evaluation database of USMLE Step 1 and Step 2 Clinical Knowledge scores, NBME subject examination scores for surgery and medicine, and surgery clerkship director ratings in the domains of clinical reasoning and fund of knowledge. Additionally, student rotation order and preceding completion of the obstetrics and gynecology or medicine clerkship were determined.

Data analysis

Data analysis was performed to generate descriptive statistics by case scenario as well as overall exam performance. Each scenario item was coded for key features of patient management, history and physical exam, formulation of a differential diagnosis or correctly identifying a final diagnosis, selection and interpretation of radiographic studies, knowledge of surgical anatomy, and communication skills. Descriptive statistics were generated for student feedback on exam factors of objectives, relevance, clarity, and feedback. Nonparametric trend analysis was performed to assess student performance on the CSE and the NBME surgery subject exam to determine seasonal variation.

Convergent validity was assessed first using bivariate analysis of existing objective measures of performance both inclusive and exclusive of the surgery clerkship: USMLE Step 1 and Step 2 Clinical Knowledge scores, NBME surgery and medicine shelf exam scores, and surgery clerkship director ratings in the domains of clinical reasoning and fund of knowledge. A multivariate linear regression model was then applied to assess the relationships of these variables and to identify which parameters were predictive of performance on the CSE and on the NBME surgery subject exam.

The effect of clerkship order or prior completion of the obstetrics and gynecology or medicine clerkship was assessed using bivariate analysis with analysis of variance for categorical variables and chi-square tests for binary variables and using multivariate analysis with a logistic regression model. To assess for the presence of construct-irrelevant variance, we looked for linear trends in score averages on cases repeated across successive exam administrations.[20,21]

Reliability of the CSE was measured using the internal consistency model, calculating a modified Cronbach's α coefficient. The standard error of measurement was calculated by multiplying exam standard deviation by the square root of the difference of 1 minus our reliability coefficient.[20] The Spearman-Brown formula was used to calculate the Cronbach's α reliability coefficient that would theoretically be achieved by increasing the number of cases on our exam.

Stata version 11.1 (StataCorp LP, College Station, TX) was used for statistical analysis, with P values $<.05$ indicating statistical significance.

Results

During the study period, 262 students representing 3 class years completed 14 unique versions of the CSE. Student performance overall was normally distributed, with a mean percentage score of 71, a median of 72, and a standard deviation of 8 (Fig. 2). Scenarios were administered to 36 to 109 students, with mean percentage scores ranging from 55 to 79, medians from 55 to 81, and standard deviations from 8 to 16. Each unique exam version was administered to 13 to 33 students, with mean percentage scores ranging from 63 to 76, medians from 55 to 77, and standard deviations from 4 to 9. The mean content distribution of the exam is shown in Fig. 3.

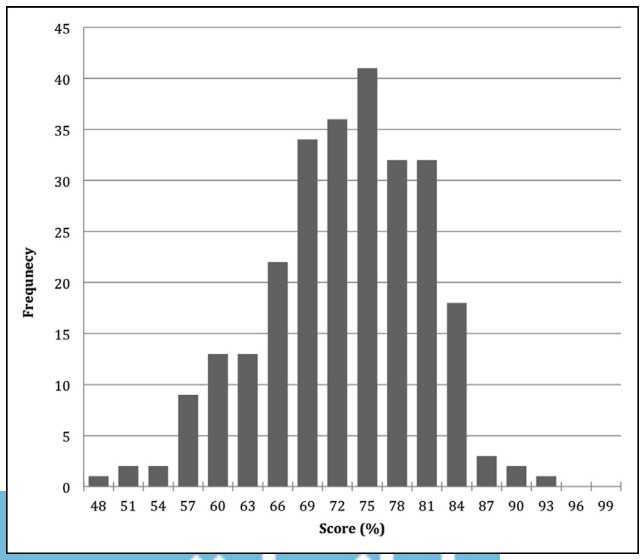

Figure 2 Distribution of student performance on the CSE.

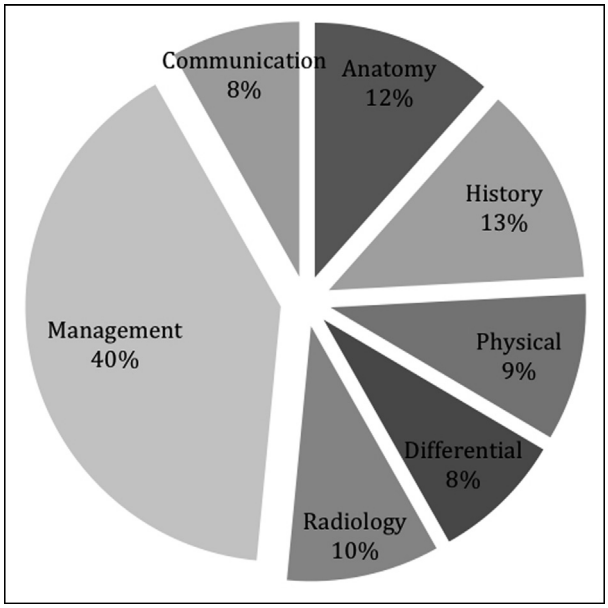

Figure 3 Descriptive statistics: CSE mean content distribution.

Nonparametric trend assessment of student performance on the CSE did not show a statistically significant variation in scores over time, in contrast with the NBME surgery subject examination, for which scores exhibited an upward trend by class year for both the raw score (Fig. 4) and the quarterly percentile score. Additionally, trend assessment on repeated scenarios across successive examination administrations confirmed the lack of a significant upward trend in score.

In bivariate analysis, CSE score correlated with all the independent measures of knowledge evaluated, except

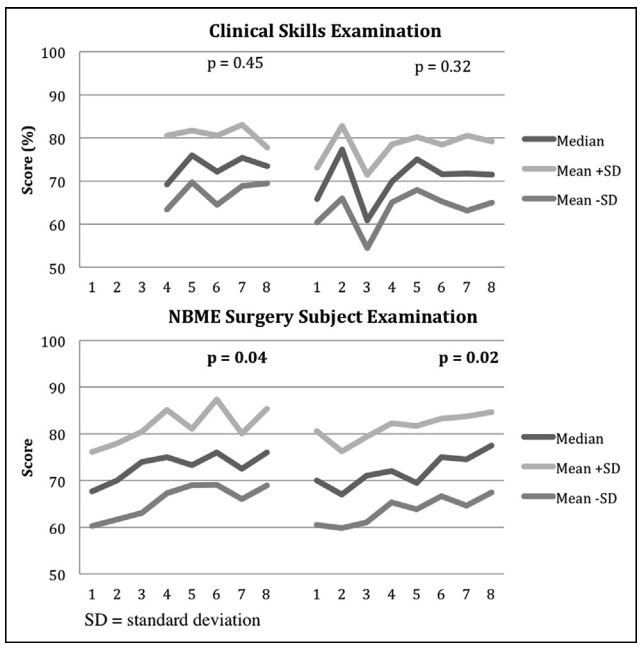

Figure 4 Trend analysis: CSE and NBME surgery subject examination by rotation group (1–8) by academic year.

Table 1 Convergent validity assessment of the CSE and NBME surgery subject examination

	CSE		NBME subject examination	
	Correlation	Linear regression	Correlation	Linear regression
	r	*P*	*r*	*P*
NBME subject examination				
Surgery	.36*	.82		
Medicine	.01	.54	.08	.20
USMLE				
Step 1	.34*	.82	.55*	.02
Step 2 Clinical Knowledge	.44*	.007	.56*	.004
Surgery clerkship composite evaluation				
Clinical reasoning	.48*	.001	.43*	.09
Fund of knowledge	.56*	.004	.63*	<.001

CSE = Surgery Clerkship Clinical Skills Examination; NBME = National Board of Medical Examiners; USMLE = United States Medical Licensing Examination.

*$P < .001$.

NBME medicine subject exam score. A multivariate linear regression model showed CSE score to be predicted by USMLE Step 2 score and clerkship rating of clinical reasoning and fund of knowledge. In comparison, NBME surgery subject exam score correlated in bivariate analysis with the same measures of knowledge and similarly did not correlate with NBME medicine shelf exam score. A multivariate linear regression analysis showed NBME surgery shelf exam score to be predicted by USMLE Step 1 and Step 2 scores and by clerkship director rating of fund of knowledge (Table 1).

In bivariate analysis, scores on the CSE were found to be independent of clerkship order, in contrast to the NBME surgery subject exam, on which performance was independently predicted by clerkship order, by having the medicine clerkship precede surgery, and by having the obstetrics and gynecology clerkship precede surgery. This association was maintained after controlling for other objective performance measures using a logistic regression analysis (Table 2).

The modified Cronbach's α coefficient for the CSE was .67. The Spearman-Brown formula predicted a theoretical Cronbach's α coefficient of .73 for exams lengthened to 6 cases. The standard error of measurement was 12.1 for the 5-scenario exam.

Table 2 Effect of clerkship order and prior clerkships on exam performance for the CSE and NBME surgery subject examination

	CSE	NBME subject examination	
	P (ANOVA/ chi-square)	*P* (ANOVA/ chi-square)	*P* (logistic regression)
Clerkship order	.30	<.001	.01
Obstetrics and gynecology prior	.34	<.001	.02
Medicine prior	.69	<.001	.03

CSE = Surgery Clerkship Clinical Skills Examination; NBME = National Board of Medical Examiners.

Student feedback on the CSE was favorable, with a mean score indicating "exceeds expectations" for all factors assessed (Fig. 5). Students' feedback score for the "overall clerkship" is included for reference.

Comments

Schwartz et al[22] argued that teaching students how to think should be a major priority of medical education in the 21st century. Multiple-choice question exams, which predominate in medical education, do not for the most part assess students' clinical reasoning or problem solving; studying for these exams therefore does not drive learning of how to think.[5,7] Multiple-choice formats have the significant limitation of assessing knowledge rather than competence built on knowledge and thus do not suitably measure Accreditation Council for Graduate Medical

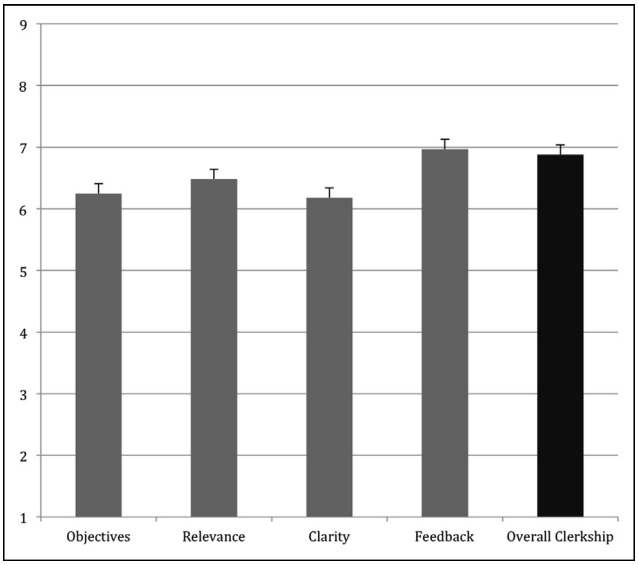

Figure 5 Student feedback on the CSE: descriptive statistics.

Education competencies beyond that of medical knowledge.[9,11,23,24] The level of Miller's pyramid most underrepresented in surgery clerkship assessment relative to its recommended weight is that of "knows how," which consists of methods designed to assess cognitive aspects of competency through the simulation of clinical problems.[12–14]

A major limitation in the development and use of methods from the "knows how" level of the pyramid has been case specificity resulting from the use of a limited number of case scenarios per exam. In designing such an exam, a balance must be achieved between test length and the psychometrics of validity and reliability. Increasing the number of cases on an exam will achieve greater validity and reliability because of decreased case specificity and examinee-by-task interaction, but a longer exam may be unfeasible or unacceptable to examinees.[5,7]

In our study, we have shown that the CSE has validity for use in the objective appraisal of students' performance on the surgery clerkship. Concurrent validity evidence demonstrates similar patterns of correlation between the CSE and external markers of student achievement that mirror those of the NBME surgery subject exam. Lack of correlation between CSE scores and NBME medicine subject exam scores constitutes divergent validity evidence that these exams measure different constructs.

Analysis of the CSE demonstrates a lack of bias in exam design. Unlike the NBME surgery subject exam, the CSE is not biased toward students completing the surgery clerkship later in the course of their clinical year or toward those students who have completed the medicine or obstetrics and gynecology clerkship before surgery. This independence from clerkship order results in added value for the CSE; it enables a fair comparison between students of different rotation groups in a way that is not possible on the basis of the NBME surgery shelf exam.

Test security appears to be adequate for the CSE, as no upward trends in average score on repeated cases were observed. Continued monitoring for such trends is necessary, with the use of the nonparametric trend test to identify upward trend significance. Elimination from the exam question database of cases demonstrating significantly upward trending scores will be necessary to protect exam validity from construct-irrelevant variance; assessing for such trends will be a continued effort within our educational program. Quarterly data analysis requires training or collaboration with a statistician or an educational specialist. This need for resource-intensive continued question development and screening to ensure validity of the CSE may be considered an exam limitation. Future efforts for cost containment to enable feasible incorporation of this or other similar exam at the national level include the development of a computer algorithm for automated grading with oversight.

Downing stated,[25] "For assessments with lower consequences, such as formative or summative classroom-type assessments, created and administered by local faculty, one might expect reliability to be in the range of .70–.79 or so." On the basis of our present analysis, the reliability of the CSE falls just short of achieving this criterion, with a reliability coefficient of .67. Our modified Cronbach's α for reliability used subtotal scores for each of the 5 scenarios of a given CSE administration; each of these subtotal scores constitutes the sum of multiple item scores. On the basis of the Spearman-Brown formula, the reliability of our exam would increase to .73 by the addition of a 6th case to each exam administration. Although the use of a longer exam would have greater psychometric reliability, we feel that this would constitute an unacceptable, unfeasible increase in exam length. The value of an exam depends on the balance of validity, reliability, and practicality, the latter contingent on feasibility, cost, acceptability.[26] At this time, the balance of these factors supports our continued use of a 5-scenario exam.

There were limitations to our statistical analysis. Interdependency of items in the linear regression model is likely, and may dampen the strength of our conclusions. The choice of parameters for our concurrent validity analysis was based on subjective judgment and available data; they may not represent the ideal construct for exam validation.

Conclusions

The CSE is an objective, valid, and reliable instrument for assessing students' performance in the surgery clerkship. Unlike the NBME surgery subject exam, performance on the CSE does not depend on clerkship order or prior clerkships, allowing the uniform assessment of students across the course of the clinical year. The CSE adds value to end-of-clerkship assessment of students' performance by providing an objective, unbiased, and seasonally independent measurement of students' achievement in areas of surgical knowledge, clinical reasoning, patient care management, and communication.

References

1. Leach D. The ACGME competencies: substance or form? J Am Coll Surgeons 2001;192:396–8.
2. Epstein RM. Assessment in medical education. New Engl J Med 2007; 356:387–96.
3. McGuire C. Perspectives in assessment. In: Gonella JS, Hojat M, Erdmann JB, et al., editors. Assessment Measures in Medical School, Residency, and Practice: The Connections. New York: Springer; 1993. p. 3–16.
4. Norcini JJ, Holmboe ES, Hawkins RE. Evaluation challenges in an era of outcomes-based education. In: Holmboe ES, Hawkins RE, editors. Practical Guide to the Evaluation of Clinical Competence. Philadelphia, PA: Mosby Elsevier; 2008. p. 1–9.
5. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ 1996;1:41–67.
6. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(suppl):S63–7.
7. van der Vleuten CPM, Newble DL. How can we test clinical reasoning? Lancet 1995;345:1032–4.

8. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ 2005;39:1188–94.
9. Schuwirth LW, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004;38:974–9.
10. Pickering G. Against multiple choice questions. Med Teach 1979;1:84–6.
11. Newble DI, Baxter A, Elmslie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. Med Educ 1979;13:263–8.
12. National Board of Medical Examiners. 2008 clinical clerkship directors survey results. Philadelphia, PA: National Board of Medical Examiners; 2008.
13. Lind DS, Deladisma AM, Cue JI, et al. Survey of student education in surgery. J Am Coll Surg 2007;204:969–74.
14. National Board of Medical Examiners. Surgery subject examination score interpretation guide. Philadelphia, PA: National Board of Medical Examiners; 2010.
15. Hawkins RE, Swanson DB. Using written examinations to assess medical knowledge and its application. In: Holmboe ES, Hawkins RE, editors. Practical Guide to the Evaluation of Clinical Competence. Philadelphia, PA: Mosby Elsevier; 2008. p. 42–59.
16. Widmann WB, Aranoff T, Fleischer BR, et al. Why should the first be last? "Seasonal" variations in the National Board of Medical Examiners (NBME) subject examination program for medical students in surgery. Curr Surg 2003;60:69–72.
17. Baciewicz Jr FA, Arent L, Weaver M, et al. Influence of clerkship structure and timing on individual student performance. Am J Surg 1990;159:265–8.
18. Ripkey DR, Case SM, Swanson DB. Predicting performance on the NBME surgery subject test and USMLE Step 2: the effects of surgery clerkship timing and length. Acad Med 1997;72:S31–3.
19. National Board of Medical Examiners. 2013–2014 subject examination fees. Available at: http://www.nbme.org/Schools/Subject-Exams/Fees_2013-2014.html. Accessed July 1, 2013.
20. Haladyna TM. Roles and importance of validity studies in test development. In: Downing SM, Haladyna TM, editors. Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum; 2006. p. 739–53.
21. Niehaus AH, DaRosa DM, Markwell SJ, et al. Is test security a concern when OSCE stations are repeated across clerkship rotations? Acad Med 1996;71:287–9.
22. Schwartz RW, Donnelly MB, Young B, et al. Undergraduate surgical education for the twenty-first century. Ann Surg 1992;216:639–47.
23. Elstein A. Beyond multiple-choice questions and essays. The need for a new way to assess clinical competence. Acad Med 1993;68:244–9.
24. Schuwirth LW, Van Der Vleuten CP, Donkers HH. A closer look at cueing effects in multiple-choice questions. Med Educ 1996;30:44–9.
25. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ 2004;38:327–33.
26. Crossley J, Haladyna TM, Jolly B. Assessing health professionals. Med Educ 2002;36:800–4.