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,*

aDepartment of Surgery, New York-Presbyterian Hospital, Columbia University Medical Center, 177 Fort Washington Avenue, MHB 7GS-313, New York, NY 10032, USA; bCenter 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. © 2014 Elsevier Inc. All rights reserved.

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
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.\(^\text{2,4,5}\)

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).\(^\text{6}\) 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.\(^\text{6,7}\) 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.\(^\text{8-10}\) Compared with multiple-choice question exams, these exams are less influenced by cueing and do not overestimate examinees’ ability.\(^\text{11}\) The major limitation of these examinations is case specificity, which results in lower reliability than may be achieved with multiple-choice question exams.\(^\text{6,7}\)

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.\(^\text{12,13}\) 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.\(^\text{12}\) 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.”\(^\text{14}\) 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.\(^\text{13}\)

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.\(^\text{15}\) 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.\(^\text{14,16-18}\) 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.\(^\text{16}\) 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.

---

**Figure 1** Miller’s pyramid framework for clinical assessment. Adapted with permission from Miller.\(^\text{6}\)
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 exam-
ination 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 succes-
sive 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. Profession-
alism 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 evalua-
tion. 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 compo-
ents of their end-of-clerkship assessments. The Depart-
ment 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 resi-
dent, faculty, and clerkship director evaluative ratings of
students’ clerkship performance as collected using software
(New Innovations, Unions Point, 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 exami-
nation scores for surgery and medicine, and surgery clerk-
ship 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 perfor-
ance. Each scenario item was coded for key features of
patient management, history and physical exam, formula-
tion of a differential diagnosis or correctly identifying a
final diagnosis, selection and interpretation of radiographic
studies, knowledge of surgical anatomy, and communica-
tion 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 clerk-
ship director ratings in the domains of clinical reasoning
and fund of knowledge. A multivariate linear
regression model was then applied to assess the relation-
ships 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. The 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. 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.

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...
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.

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. 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. 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

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

<table>
<thead>
<tr>
<th></th>
<th>CSE Correlation</th>
<th>CSE Linear regression</th>
<th>NBME subject examination Correlation</th>
<th>NBME subject examination Linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery</td>
<td>.36*</td>
<td>.82</td>
<td>.08</td>
<td>.20</td>
</tr>
<tr>
<td>Medicine</td>
<td>.01</td>
<td>.54</td>
<td>.08</td>
<td>.20</td>
</tr>
<tr>
<td>USMLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td>.34*</td>
<td>.82</td>
<td>.55*</td>
<td>.02</td>
</tr>
<tr>
<td>Step 2 Clinical Knowledge</td>
<td>.44*</td>
<td>.007</td>
<td>.56*</td>
<td>.004</td>
</tr>
<tr>
<td>Surgery clerkship composite evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical reasoning</td>
<td>.48*</td>
<td>.001</td>
<td>.43*</td>
<td>.09</td>
</tr>
<tr>
<td>Fund of knowledge</td>
<td>.56*</td>
<td>.004</td>
<td>.63*</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

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

*P < .001.

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

<table>
<thead>
<tr>
<th></th>
<th>CSE P (ANOVA/chi-square)</th>
<th>NBME subject examination P (ANOVA/chi-square)</th>
<th>P (logistic regression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clerkship order</td>
<td>.30</td>
<td>&lt;.001</td>
<td>.01</td>
</tr>
<tr>
<td>Obstetrics and gynecology prior</td>
<td>.34</td>
<td>&lt;.001</td>
<td>.02</td>
</tr>
<tr>
<td>Medicine prior</td>
<td>.69</td>
<td>&lt;.001</td>
<td>.03</td>
</tr>
</tbody>
</table>

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

### Figure 5 Student feedback on the CSE: descriptive statistics.
Education competencies beyond that of medical knowledge. 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. 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.

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: “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 a 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. 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