Association for Surgical Education

Faculty and resident evaluations of medical students on a surgery clerkship correlate poorly with standardized exam scores

Seth D. Goldstein, M.D. a,*, Brenessa Lindeman, M.D. a, Jorie Colbert-Getz, Ph.D. b, Trisha Arbella, B.S. a, Robert Dudas, M.D. c, Anne Lidor, M.D. a, Bethany Sacks, M.D. a

aDepartment of Surgery, Johns Hopkins School of Medicine, 1800 Orleans Street, Bloomberg Children’s Center 7310, Baltimore, MD 21287, USA; bOffice of Medical Education Services, Johns Hopkins School of Medicine, Baltimore, MD, USA; cDepartment of Pediatrics, Johns Hopkins School of Medicine, Baltimore, MD, USA

KEYWORDS: Medical student education; Assessment; Surgery clerkship

Abstract

BACKGROUND: The clinical knowledge of medical students on a surgery clerkship is routinely assessed via subjective evaluations from faculty members and residents. Interpretation of these ratings should ideally be valid and reliable. However, prior literature has questioned the correlation between subjective and objective components when assessing students’ clinical knowledge.

METHODS: Retrospective cross-sectional data were collected from medical student records at The Johns Hopkins University School of Medicine from July 2009 through June 2011. Surgical faculty members and residents rated students’ clinical knowledge on a 5-point, Likert-type scale. Interrater reliability was assessed using intraclass correlation coefficients for students with ≥4 attending surgeon evaluations (n = 216) and ≥4 resident evaluations (n = 207). Convergent validity was assessed by correlating average evaluation ratings with scores on the National Board of Medical Examiners (NBME) clinical subject examination for surgery. Average resident and attending surgeon ratings were also compared by NBME quartile using analysis of variance.

RESULTS: There were high degrees of reliability for resident ratings (intraclass correlation coefficient, .81) and attending surgeon ratings (intraclass correlation coefficient, .76). Resident and attending surgeon ratings shared a moderate degree of variance (19%). However, average resident ratings and average attending surgeon ratings shared a small degree of variance with NBME surgery examination scores ($r^2 \leq .09$). When ratings were compared among NBME quartile groups, the only significant difference was for residents’ ratings of students with the lower 25th percentile of scores compared with the top 25th percentile of scores ($P = .007$).

CONCLUSIONS: Although high interrater reliability suggests that attending surgeons and residents rate students with consistency, the lack of convergent validity suggests that these ratings may not be reflective of actual clinical knowledge. Both faculty members and residents may benefit from training in knowledge assessment, which will likely increase opportunities to recognize deficiencies and make student evaluation a more valuable tool.

© 2014 Elsevier Inc. All rights reserved.

The authors declare no conflicts of interest.

* Corresponding author. Tel.: +1-410-955-2717; fax: +1-410-502-5314.
E-mail address: sgoldstein@jhu.edu

Manuscript received May 6, 2013; revised manuscript October 2, 2013

0002-9610/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amjsurg.2013.10.008
Fostering the development of clinical knowledge is among the primary goals of medical student clerkships, but no gold standard for assessment has emerged in this key area. Common approaches to knowledge assessment on clinical clerkships at most medical schools remain a mixture of subjective evaluations from faculty members and residents with objective scores on national standardized examinations. Student assessment should ideally be valid and reliable; however, prior literature has demonstrated mixed conclusions when examining correlations between subjective and objective components of student clinical knowledge. Literature from radiology and pediatrics has demonstrated moderate correlations between grades from subjective and objective components of medical knowledge. However, other studies in emergency medicine and internal medicine have shown lower levels of correlation between medical knowledge assessment by faculty members and discipline-specific standardized exam performance.

Only 1 prior study has also examined evaluations of surgical students, demonstrating low predictive value of resident ratings that was only marginally better than the predictive value of surgical faculty member ratings. These points are of key importance not only regarding the administrative decision of what grades to assign students but also because early recognition of deficits in student performance is crucial in offering constructive strategies to overcome them. Although self-assessment is a key component of adult learning, research has repeatedly demonstrated poor correlations between medical students’ self-assessments with objective measures of knowledge and their final clerkship grades.

Rigorous validation of scores from subjective assessments on student clerkships has not been conducted, although all medical schools in the United States use these in the clinical years. One study showed that a student’s overall assigned clerkship grade can be predicted by faculty ratings in only a single performance area, despite these ratings’ not correlating with standardized, objective measures. Because of the potential predictive ability of subjective ratings, instructors who sense deficiencies in students’ performance are able to provide timely feedback and work with learners to adapt learning plans earlier during the clerkship.

This study was designed to investigate the convergent validity between subjective ratings of clinical knowledge and scores on the National Board of Medical Examiners (NBME) subject examination, as well as interrater reliability of faculty members’ and residents’ evaluations of global clinical knowledge among students on the surgery clerkship. We hypothesized that surgical residents’ and faculty members’ ratings of clinical knowledge would correlate poorly with the students’ standardized exam scores.

### Methods

Retrospective cross-sectional data were collected from medical student records at The Johns Hopkins University School of Medicine from July 2009 through June 2011 (n = 219 students ranging from the 2nd to the 4th year). The medical student basic clerkship was just under 9 weeks in duration and was divided into a 4.5-week general surgery experience and 2 separate 2-week surgical subspecialty rotations, though not necessarily in that order. Students were instructed to approach potential evaluators at the conclusion of their time on a given service to request evaluations, which were then sent by e-mail and completed within 4 weeks. Minimums of 4 faculty member and 4 resident evaluations were desired. Surgical faculty members and residents rated students’ clinical knowledge as part of a 17-item summative evaluation. All items were rated on a 5-point, Likert-type scale. The clinical knowledge 1-to-5 rating descriptors are provided in Table 1.

Data analysis was performed using SPSS version 20 (IBM, Armonk, NY). The clinical knowledge rating was extracted from the full evaluation, and the interrater reliability, or consensus between evaluators, of those scores was assessed. The proportion of variance due to variability of scores between raters, known as the intraclass correlation coefficient (ICC), was calculated separately for both faculty member and resident ratings. An ICC $\geq .75$ indicates good agreement among raters and thus good reliability. Values of $.50$ to $.74$ indicate moderate reliability, and values $<.49$ indicate poor reliability.

Convergent validity of clinical knowledge ratings was assessed by correlating average ratings with scores on the NBME clinical subject examination for surgery using Spearman’s $\rho$. The $\rho^2$ value was also calculated to determine the shared variance between ratings and examination scores. A $\rho^2$ value $\geq .25$ indicates a high degree of

<table>
<thead>
<tr>
<th>Table 1 Clinical knowledge rating scale (5-point, Likert-type scale)</th>
</tr>
</thead>
</table>
| 1 Unacceptable
   Unable to apply preclinical knowledge to understand basic medical problems. |
| 2 Needs improvement
   Inconsistent understanding of patient problems.
   Limited differential diagnosis. |
| 3 At expected level
   Knows basic differential diagnoses of major/active problems in patients.
   Understands team’s choice of therapy. |
| 4 Above expectations
   Knows expanded differential diagnoses, including recognition of emergencies.
   Can discuss therapeutic options. |
| 5 Outstanding
   Knows nuances of differential diagnosis, including disease prevalence and anticipated history and exam findings. Able to independently formulate a management plan. Able to assign prognoses. |
variance shared between 2 variables, values of .09 to .24 indicate a moderate degree of variance, and values <.09 indicate a small degree of variance. Additionally, students were assigned to 1 of 3 groups on the basis of their examination scores relative to their peers’ scores: top 25%, middle 50%, and bottom 25%. Average clinical knowledge ratings from residents and attending surgeons were analyzed for differences on the basis of NBME score quartile using analyses of variance.

Results

The median and mode number of clinical knowledge ratings for each student were 4 from attending surgeons and 4 from residents. To calculate the ICC, each student must be rated by the same number of raters. Thus, students who had <4 evaluations from residents or <4 evaluations from attending surgeons were omitted from further analyses. There were 216 students with ≥4 attending surgeon evaluations and 207 students with ≥4 resident evaluations. A random sample of 4 attending surgeon evaluations and 4 resident evaluations was selected for students with >4 evaluations for either rater type. The frequency of ratings given was as follows: by attending surgeons, no 1’s, 4 (.5%) 2’s, 178 (20%) 3’s, 498 (57%) 4’s, and 193 (22%) 5’s; by residents, no 1’s, 6 (.7%) 2’s, 244 (28%) 3’s, 402 (47%) 4’s, and 211 (24%) 5’s. The ICC for residents was .81, and the ICC for attending surgeons was .76, suggesting a high degree of agreement in ratings.

The correlation coefficients (p values) between average ratings and surgical examination scores were .20 for residents and .12 for attending surgeons. NBME examination scores shared a small degree of variance with average resident ratings and average attending surgeon ratings, 1% and 1%, respectively. Fig. 1 shows the 95% confidence intervals for average resident and attending surgeon ratings by NBME examination group. Residents had average ratings of 3.81 ± .48 for the bottom 25th group, 3.95 ± .43 for the middle 50th group, and 4.10 ± .49 for the top 25th group. These ratings were significantly different, \( F(2, 204) = 5.12, P = .007, \eta^2_p = .05 \). Post hoc analysis revealed that the difference between the top 25th group and the bottom 25th group was driving the effect (\( P = .007 \)), as there was no significant difference between the middle 50th group and the other groups. Attending surgeons had average ratings of 3.96 ± .45 for the bottom 25th group, 4.00 ± .35 for the middle 50th group, and 4.08 ± .40 for the top 25th group. These ratings were not significantly different, \( F(2, 213) = 1.32, P = .270 \).

Comments

Assessment of residents and medical students in the United States has become increasingly aligned with the model of 6 core competencies developed by the Accreditation Council for Graduate Medical Education. Medical knowledge is 1 of the fundamental domains in which practicing physicians are called to exhibit competence. Although it is debatably the easiest to measure, there is no present consensus on which assessment methods should constitute a gold standard. As such, multiple studies have shown that wide variability exists in grading schema for medical student clerkships, as medical schools have each applied their own standards on an ad hoc basis.

Our findings suggest that although the magic bullet of medical student performance assessment continues to remain elusive, there may be predictable patterns. Although the convergent validity of subjective assessments from both faculty members and residents with NBME surgery subject examination scores was low, the interrater reliability between multiple faculty member and resident ratings for each student was high. On that basis, we posit that other aspects of student performance are perhaps being considered as proxies for knowledge. Data from a prior study that examined factors contributing to faculty members’ evaluation of medical students’ performance on a surgery clerkship indicated that faculty members form 1-dimensional views of students’ performance when assigning grades, rather than nuanced cognitive models that account for differentiation among multiple grading criteria. Our data further imply that faculty evaluators in surgery may be conceptualizing a generalized global assessment of student performance on which they base all of their ratings. This is colloquially known as the halo effect.

Interestingly, we found that residents’ evaluations of students’ knowledge correlated better with NBME examination scores than did faculty members’ evaluations.
Studies have shown that students are observed infrequently by faculty members in clinical encounters and spend a majority of their contact time with residents. There is documentation of variation in faculty members’ evaluations of students by clinical service, with increasing length of rotation (4 vs 2 weeks) correlating with lower overall scores, although no similar data exist for residents’ evaluations. We agree with Dudas et al that “fragmentation of student clinical experiences is a threat to assessment,” and it can be reasonably argued that longitudinal clerkship experiences provide the maximal opportunities for accurate student assessment across all domains of competency.

Residents’ ability to statistically differentiate students in the top quartile of exam scores from those in the bottom quartile could be due to a combination of quantitative and qualitative differences in time spent with clerkship students. The change in ratings with extent of exposure may be explained by the provision of additional opportunity for cognitive and interpersonal skills evaluation over the course of rounds, time in the operating room, and clinics. Additionally, residents may simply be offering more nuanced assessments because of their proximity to students’ age, social dynamic, and position in the learning hierarchy. Older attending surgeons may benefit from guidance regarding the expectations of student learning progression, as Tortolani et al demonstrated that the average faculty member’s expectation for fulfilling learning objectives at the end of a surgical clerkship differs only minimally from expectations at the beginning of the clerkship. Such data from other specialties are limited.

Another possible explanation for the discrepancy between objective and subjective ratings of medical student clinical knowledge in surgery may lie in the raters’ interpretations of what they are being asked to score: simple factual knowledge versus a student’s ability to apply that knowledge. Evaluators may very well be judging students’ ability to critically apply their clinical knowledge, a task for which a higher cognitive load is required, rather than the largely factual recall required for the NBME examination. Thus, variation in the perceived construct measured by the question may result in the lack of convergent validity observed between student clinical knowledge assessments and examination scores.

Although our results are in concordance with those of similar studies conducted at other institutions, relying on data from a single institution is a limitation of this study and may affect the generalizability of our findings. Additionally, the duration and structure of the surgical clerkship experiences included vary somewhat among specialties and were incompletely documented. These factors are likely to affect faculty members’ contact time with students and the numbers of faculty members and residents who evaluated each student. More research is necessary to further delineate factors that affect the validity of subjective ratings in medical student clerkships.

Conclusions

Assessment of medical students is of broad interest and should ideally be valid and reliable. Medical schools use both subjective and objective assessments of performance on student clerkships, as each contributes differently to insight regarding students’ abilities. However, it is unknown to what extent clinicians can detect excellence or deficiencies while working with medical students. Our data suggest that there are contextual and discipline-specific trends at work regarding the ways in which faculty members and residents perceive and subsequently rate students’ performance. Specifically, compared with medical specialties, the surgical routines and culture may not easily lend themselves to accurate student assessment without focused training in such. We have yet to explore if and how the importance given to technical skills in surgery skews our ability to assess other domains compared with nonprocedural disciplines. Within the confines of the existing student clerkship structure, it seems evident that both faculty members and residents would benefit from dedicated training in rating students’ knowledge. Such an intervention could decrease the “halo effect” that has been demonstrated to plague the subjective evaluation of medical students and would also provide increased opportunities to recognize deficiencies within the context of the time-limited settings so common in surgical practice.

References