The American
Journal of Surgery®

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]

[a]Department of Surgery, Johns Hopkins School of Medicine, 1800 Orleans Street, Bloomberg Children's Center 7310, Baltimore, MD 21287, USA; [b]Office of Medical Education Services, Johns Hopkins School of Medicine, Baltimore, MD, USA; [c]Department of Pediatrics, Johns Hopkins School of Medicine, Baltimore, MD, USA

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 ($\rho^2 \leq .09$). 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.

The authors declare no conflicts of interest.
* Corresponding author. Tel.: +1-410-955-2717; fax: +1-410-502-5314.
E-mail address: sgoldstein@jhmi.edu

Fostering the development of clinical knowledge is among the primary goals of medical student clerkships,[1] 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.[2,3] 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.[4–6] Only 1 prior study has also examined evaluations of surgical students,[7] 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[8,9] and their final clerkship grades.[10]

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.[11] One study showed that a student's overall assigned clerkship grade can be predicted by faculty ratings in only a single performance area,[12] 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.[13]

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 ρ. The ρ^2 value was also calculated to determine the shared variance between ratings and examination scores. A ρ^2 value $\geq$ .25 indicates a high degree of

Table 1 Clinical knowledge rating scale (5-point, Likert-type scale)

1	2	3	4	5
Unacceptable	Needs improvement	At expected level	Above expectations	Outstanding
Unable to apply preclinical knowledge to understand basic medical problems.	Inconsistent understanding of patient problems. Limited differential diagnosis.	Knows basic differential diagnoses of major/active problems in patients. Understands team's choice of therapy.	Knows expanded differential diagnoses, including recognition of emergencies. Can discuss therapeutic options.	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.