Is a Single-Item Operative Performance Rating Sufficient?

Reed G. Williams, PhD,* Steven Verhulst, PhD,† John D. Mellinger, MD,‡ and Gary L. Dunnington, MD*

*Department of Surgery, Indiana University School of Medicine, Indianapolis, Indiana; †Department of Medical Education, Southern Illinois University School of Medicine, Springfield, Illinois; and ‡Department of Surgery, Southern Illinois University School of Medicine, Springfield, Illinois

OBJECTIVE: A valid measure of resident operative performance ability requires direct observation and accurate rating of multiple resident performances under the normal range of operating conditions. The challenge is to create an operative performance rating (OPR) system that: is easy to use, encourages completion of many ratings immediately after performances and minimally disrupts supervising surgeons' work days. The purpose of this study was to determine whether a score based on a single-item overall OPR provides a valid and stable appraisal of resident operative performances.

DESIGN: A retrospective comparison of a single-item OPR with a gold-standard rating based on multiple procedure-specific and general OPR items.

SETTING: Data were collected in the general surgery residency program at Southern Illinois University from 2001 through 2012.

PARTICIPANTS: Assessments of 1033 operative performances (3 common procedures, 2 laparoscopic, and 1 open) by general surgery residents were collected. OPRs based on single-item overall performance scale scores were compared with gold-standard ratings for the same performances.

RESULTS: Differences in performance scores using the 2 scales averaged 0.02 points (5-point scale). Correlations of the single-item and gold-standard scale scores averaged 0.95. Based on generalizability analyses of laparoscopic cholecystectomy ratings, each instrument required 5 observations to achieve reliabilities of 0.80 and 11 observations to achieve reliabilities of 0.90. Only 4.4% of single-item ratings misclassified the performance when compared with the gold-standard rating and all misclassifications were near misses. For 80% of misclassified ratings, single-item ratings were lower.

CONCLUSIONS: Single-item operative performance measures produced ratings that were virtually identical to gold-standard scale ratings. Misclassifications occurred infrequently and were minor in magnitude. Ratings using the single-item scale: take less time to complete, should increase the sample of procedures rated, and encourage attending surgeons to complete ratings immediately after observing performances. Face-to-face and written comments and suggestions should continue to be used to provide the granular feedback residents need to improve subsequent performances. (J Surg 72:e212-e217. © 2015 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: operative performance evaluation, surgical education, resident training, general surgery

COMPETENCIES: Patient Care, Medical Knowledge, Practice-Based Learning and Improvement

When it comes to workplace-based assessment in the United States, surgery residency program faculty have an advantage over counterparts in other medical training environments as an attending surgeon observes every resident performance in the operating room owing to reimbursement regulations. Direct observation is the first critical condition required for valid assessment of any performance. A second critical condition in rating single performances is that performance ratings must be completed promptly after observation while the performance is fresh in the mind of the expert judge. Obviously, the ideal situation is one where the performance rating is completed as the performance occurs or immediately after observing the performance. Williams et al.[1] studied the effect of delays in completing operative performance (OPR) rating forms on the clarity and detail of the ratings recorded. They found that: ratings completed immediately possess the most clarity and detail, ratings completed within 72 hours have reasonable clarity and detail, and ratings completed more than 14 days after the performance have very poor clarity and detail. Kim et al.[2] documented that the average OPR in field conditions is completed 11 days after observing the

Correspondence: Inquiries to Reed G. Williams, PhD, Department of Surgery, Indiana University School of Medicine, Mailing Address: 108 Maple Grove, Springfield, IL 62712; E-mail: reedwill@iupui.edu, reedsuewms@comcast.net

  1931-7204/$30.00 http://dx.doi.org/10.1016/j.jsurg.2015.05.002

performance. Therefore, the challenge for residency programs is to arrange conditions that make it easy for expert raters to complete ratings more promptly if not immediately after observation of a resident operative performance.

Commonly used OPR forms consist of approximately 10 items leading some surgeons to resist completing the evaluations immediately after observation. Shorter forms have the potential to increase the total number of rated performances and may increase the number submitted within the recommended 72 hours. Ostensibly, increasing the number of performances rated should also increase the range of operative conditions (e.g., procedures performed, case complexity, and operative team composition) under which the performances occurred.

The current study is designed to determine the effect of reducing OPR instruments to a single overall performance item. This change would reduce the amount of time required to complete the form and would perhaps entice a greater percentage of surgeons to complete the form immediately after observing the performance. But the question remains, does the value added in the form of increased quantity of ratings outweigh the cost of information lost.

The specific questions to be answered in this study are:

1. What is the difference in average operative performance scores assigned when using a single overall performance item rather than a multi-item procedure-specific OPR scale?
2. How similar are the rankings of operative performances when rated using a single overall performance item and when using a multi-item OPR scale?
3. Accepting the score from the multi-item scale as the best available indication of that resident's performance on that day under that set of circumstances, what number and percentage of performances would be misclassified using the single-item operative performance scale instead? How many times would the single-item operative performance scale underestimate/overestimate the quality of a performance? How large would the differences be?
4. How many observations are required to achieve a stable estimate of operative performance using the single overall performance item vs using the multi-item scale? Presumably the shorter form would result in larger numbers of observations for each resident. Would this increase in number of observations offset the reliability lost when reducing the number of items on the scale?

METHODS

Setting

Data for this study were collected in the General Surgery Residency Program at Southern Illinois University School of Medicine from 2001 to 2012. The data included faculty ratings of 1033 operative performances (455 laparoscopic cholecystectomies, 315 laparoscopic appendectomies, and 263 open inguinal hernia repairs) by general surgery residents at all levels of training and were collected in the normal course of educating and evaluating these residents. The local institutional review board for protection of human subjects reviewed the protocol for this research project and judged that it was exempt from continuing review. The study was conducted according to the protocol submitted.

Performance Rating Instruments

The OPR system used was developed at Southern Illinois University for evaluating resident performance of key operative procedures. These OPR instruments have been refined based on evidence-based parameters of good and poor patient outcomes.[2-4] Each operative performance instrument contains 3 to 5 procedure-specific items and 5 general items appropriate for rating any operative procedure. Of these, 4 general items were developed and validated at the University of Toronto.[5] The fifth was an overall performance item. All items used 5-point Likert scales (1 = poor, 2 = fair, 3 = good, 4 = very good, and 5 = excellent) and all but the overall performance item had behavioral anchors at the 1, 3, and 5 positions on the scale. The operative procedures investigated in this study as outlined above are performed frequently by general surgery residents and by general surgeons in practice.[6] All OPR instruments are available for inspection online.[7]

Rating Process

The participating surgeons observed the operative performance while they supervised and assisted the performing resident. At a later time of their choosing the participating surgeons completed the OPR form.

Data Analysis

The multi-item scale score was formed by computing a mean total operative performance score for each performance rated. This was done by summing all performance quality item ratings (ratings for all items except the case difficulty item and the guidance item) and dividing by the number of performance quality rating items on the scale. This included the overall performance item. The single-item scale score was the rater response to the overall performance item alone.

A 1-way analysis of variance was used to determine differences in average ratings for all performances using the multi-item scale score and using the single-item scale score as the performance quality indices. Differences that