The American Journal of Surgery 215 (2018) 395–398

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

Accurate surgical skills evaluation: Does it mandate raters have a medical background?

Muhammad Zeb [a], Apram Jyot [a], Eduardo Abbott [b], Miguel Gomez [a], Mohamed Baloul [a], David Farley [a, b, *]

[a] Department of Surgery, Mayo Clinic College of Medicine, Rochester, MN, USA
[b] Multidisciplinary Simulation Center, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Article history:
Received 7 July 2017
Received in revised form
1 November 2017
Accepted 1 November 2017

Keywords:
Assessment
Raters
Evaluations
Surgical skills

ABSTRACT

Background: Surgeons rarely have time to assess/rate trainee performance. From a 10 year-experience of implementing OSCE style assessments, we hypothesize that the accurate scoring of interns in selected tests is not affected by the rater's medical background.

Methods: A prospective collection of quantitative scoring data by both medical school graduates and college students was compared. Each rater underwent training and then watched three videos of actors performing in each of two OSCE stations and individually scored them.

Results: Twelve college students and 16 medical graduates participated. There was no difference in the mean scores between rating groups for chest tube insertion (Video 1: 1.7 vs. 2.0; Video 2: 2.9 vs 3.1; Video 3: 6.1 vs 6.1; $p = 0.8$) and cricothyrotomy (Video 1: 4.0 vs 4.5; Video 2: 4.8 vs 5.1; Video 3: 9.2 vs 9.1; $p = 0.7$).

Conclusion: Accurate scoring of surgical performance does not mandate a medical background. Given the limited availability of attending surgeons for assessments, use of validated, simple checklists can help raters with minimal medical experience perform assessments proficiently.

1. Introduction

Frequent assessment and voluminous feedback are important factors within any high-powered educational system.[1] Assessment is integral to determining clinical competence and readiness to practice in medical or surgical professions.[2] Assessment allows examiners to make either an objective or subjective judgment as to whether a student has achieved competency. Ultimately, positive assessments allow medical students to become residents, and residents to become practicing physicians – capable of autonomously caring for their patients.[3] Thus, accurate assessment is vital and invariably depends strictly on human judgment and analysis to determine the outcome.[4] Human beings are known to harbor bias, often fail to document serious deficits, and may be unprepared to evaluate the learners they assess. Compounding the struggle for accurate evaluation, added potential sources of inaccuracy relate to the mechanics of the rating task, the system used to obtain ratings, and factors affecting rater judgment.[5]

In an ideal world, surgical trainees would undergo intense surgical education and training interspersed with accurate feedback, frequent assessments, and then individualized remediation and deliberate practice from expert surgeons. Additionally, assessments would follow to confirm competency, proficiency, or ideally, mastery. While we believe general surgery residents are engaged in laudable surgical education, it is especially difficult to offer an accurate assessment that is efficient, timely, and inexpensive. Staff surgeons are pressured to generate clinical productivity (RVUs), partake in research efforts, perform administrative duties, and manage the electronic medical record. There is little time for teaching residents, and less time to formally assess them.

The Mayo Clinic-Rochester general surgery program offers intense surgical education, believes in frequent assessments, and offers individualized remediation. However, we struggle with surgical staff input into the betterment of our trainees. Nonetheless, we have pushed forward with assessment efforts and have relied on a variety of sources for raters to help us accurately assess our

* Corresponding author. Department of Surgery, Mayo Clinic, 200 First Street SW, Rochester, MN, 55905, USA.
E-mail address: farley.david@mayo.edu (D. Farley).

https://doi.org/10.1016/j.amjsurg.2017.11.012
0002-9610/

trainees. Medical graduates (MD, MBBS, DO) doing research at Mayo Clinic gravitate to our Simulation Center for learning opportunities – and we subsequently put many of them to work helping us. College undergraduates (participating in research, education, and observership opportunities) on our campus each summer (July) and over their winter break (January) were added to our list of helpers to assess our trainees.

Our program conducts biannual assessments of first-year surgical residents (interns) properly called the Surgical Hexadecathlon – fondly termed the "Surgical Olympics". The assessment is divided into eight technical skills and eight knowledge based OSCE style stations that are assessed by raters using objective scoring checklists. The scoring sheets are simple, objective (typically binary - "yes" or "no") and efficient (most scores are complete by the time the trainee walks out of the testing room).

Given we utilize 5–10 raters to assist us with our Surgical Hexadecathlon, we have been nervous that such inexperienced "surgeon raters" may affect the accuracy of our assessments. We hypothesized that a brief training session to enthusiastic college students would lead them to be as accurate and as efficient as our medical school graduates. As the role of the examiner is central in performance assessment, this research sought to determine the role of medical background of examiners in evaluating surgical skills.

2. Methods

Our study was a part of an initiative for research in simulation education, approved by the Institutional Review Board at Mayo Clinic. All study participants provided consent before matriculation.

We conducted a prospective collection of quantitative scoring data by medical graduates and college students. Medical graduates were defined as the raters having an MD degree or an equivalent (i.e., MBBS, DO). The college student group included college undergraduates and medical students who were in their 1st month of medical school. Limiting our medical school participants to the first month was a conscious decision since at this point in their curriculum these medical students had not taken anatomy classes.

Cricothyrotomy and chest tube insertion were chosen as the OSCE skills to be evaluated. These skills were chosen as their scoring sheets have remained consistent for 10 years; the chest tube insertion checklist is a validated tool, and the cricothyrotomy scoring checklist was generated with a Delphi analysis utilizing 14 general and trauma surgeons. Initially a pilot study was conducted with five college students as new raters. Based on that limited effort and feedback from the prospective raters, we made changes to the study design, the instructional video, and the practice time allotted with the models.

The rater training process was divided into two major parts: 1) Instructional video and 2) Performance of the skills by the raters. Initially the raters were shown an instructional (10 min) video demonstrating the basic anatomy relevant to chest tube insertion and cricothyrotomy, instruments used in the task, description of the models (Fig. 1), and the scoring sheets. Raters were allowed to pause or replay the video. During the video, raters had access to the models, the instruments and the individual scoring sheets so they could have a better understanding of the tools used in the video. After the video was shown, representation of different anatomical structures on the model were explained to the raters in person and were given a chance to ask questions. They were then asked to perform the two respective skills; they were scored on their performance. Feedback was provided according to their performance and how it transferred to the score sheet/checklist. The overall process took approximately 30 min to complete.

Medical students and college students were then asked to score three videos of trainees performing chest tube insertion and three performing cricothyrotomy. Each video depicted a different level of performance; the videos were played in a random order to prevent confirmation bias. Perfect chest tube insertion was given a total score of seven; the maximum score for cricothyrotomy was 10.

The primary outcome was interrater reliability. Secondary outcomes included score accuracy and the mean scores. Cohen's kappa statistic was used to compute interrater reliability (IRR). IRR statistics for nominal data were formalized as extensions of Scott's (1955) Pi statistic (e.g., Fleiss's 1971). Means were calculated using a two-sided alpha of 0.05. The analysis was conducted using JMP software (2012 SAS Institute Inc).

3. Results

There were 16 raters (mean age = 27 years, range: 25–31) in the medical graduate group and 12 in the college student group (mean age = 22 years, range: 20–25). Of 15 medical graduates, 14 were research fellows working towards residency acceptance and one was a pediatric resident. Of 12 college students, 8 were in their first month of the 1st year of medical school at Mayo Clinic and 4 were college undergraduate students. There was no difference in the mean scores between the two groups: Cricothyrotomy - college student rater mean score = 6.0 ± 2.3 vs medical graduate rater score = 6.1 ± 2.2 ($p = 0.7$); Chest tube insertion - college student rater mean score = 3.6 ± 1.9 vs medical graduate rater mean score = 3.7 ± 1.9 ($p = 0.8$). All six videos were consistently rated by each group ($p > 0.05$; Table 1) and matched closely the scoring of experienced raters and the senior surgeon-author. Mean kappa for chest tube and cricothyrotomy for college students was 0.68 and for medical graduates was 0.65, respectively (Substantial agreement per Landis and Koch[6]).

4. Discussion

This simplistic study suggests that non-medical personnel can: 1) accurately evaluate examinee performance recorded on video, 2) learn how to rate skill performance through a limited hands-on and video instruction curriculum, 3) accurately apply the use of a simplistic checklist, and 4) become potentially of great use to surgeon educators looking to offer their trainees voluminous and accurate feedback.

A study conducted on neurosurgery residents showed that assessment of video recorded technical skills of residents can be measured with high interrater reliability and surgeons and non-surgeons alike readily distinguish different skill levels although the study did not define the educational level of non-surgeons.[7] Another study focused on the educational level of raters showed that the undergraduate students were more accurate and consistent in their ratings.[8] On the other hand, another study confirmed that the rater's background did not affect the ratings and that it was the student's actual performance on a task that contributed the greatest variance in ratings.[9] Additionally, video recordings can be used to coach residents, to track performance over time, and to compare skill levels.

It is important to note that the skills, chest tube insertion and cricothyrotomy used checklists. The chest tube scoring checklist is validated and the cricothyrotomy scoring checklist has been developed by consensus of multiple experts, hence providing minimal discrepancy while scoring. A systematic review showed that the average inter-rater reliability was high and slightly better