A multi-institutional study of the Family Conference Objective Structured Clinical Exam: a reliable assessment of professional communication

Jeffrey G. Chipman, M.D.\textsuperscript{a,*}, Travis P. Webb, M.D.\textsuperscript{b}, Mohsen Shabahang, M.D., Ph.D.\textsuperscript{c}, Stephanie F. Heller, M.D.\textsuperscript{d}, Joan M. van Camp, M.D.\textsuperscript{e}, Amy L. Waer, M.D.\textsuperscript{f}, Michael G. Luxenberg, Ph.D.\textsuperscript{g}, Matthew Christenson\textsuperscript{g}, Connie C. Schmitz, Ph.D.\textsuperscript{a}

\textsuperscript{a}Department of Surgery, University of Minnesota, 420 Delaware St S.E., 11-157 Phillips Wangensteen Building, MMC 190, Minneapolis, MN 55455; \textsuperscript{b}Department of Surgery, Medical College of Wisconsin, Milwaukee, WI; \textsuperscript{c}Texas A&M, Scott and White Clinic, Temple, TX; \textsuperscript{d}Mayo Clinic, Rochester, MN; \textsuperscript{e}Hennepin County Medical Center, Minneapolis, MN; \textsuperscript{f}Department of Surgery, University of Arizona, Tucson, AZ; \textsuperscript{g}Professional Data Analysts, Inc, Minneapolis, MN, USA

Abstract

\textbf{BACKGROUND:} To test the value of a simulated Family Conference Objective Structured Clinical Exam (OSCE) for resident assessment purposes, we examined the generalizability and construct validity of its scores in a multi-institutional study.

\textbf{METHODS:} Thirty-four first-year (PG1) and 27 third-year (PG3) surgery residents (n = 61) from 6 training programs were tested. The OSCE consisted of 2 cases (End-of-Life [EOL] and Disclosure of Complications [DOC]). At each program, 2 clinicians and 2 standardized family members rated residents using case-specific tools. Performance was measured as the percentage of possible score obtained. We examined the generalizability of scores for each case separately. To assess construct validity, we compared PG1 with PG3 performance using repeated measures multivariate analysis of variance (MANOVA).

\textbf{RESULTS:} The relative G-coefficient for EOL was .890. For DOC, the relative G-coefficient was .716. There were no significant performance differences between PG1 and PG3 residents.

\textbf{CONCLUSIONS:} This OSCE provides reliable assessments suitable for formative evaluation of residents’ interpersonal communication skills and professionalism.

© 2010 Elsevier Inc. All rights reserved.

\textbf{KEYWORDS:} Simulation; Objective Structured Clinical Exam (OSCE); Surgical resident; Assessment; End-of-life; Complications; Communication

This study was funded by the Center for Excellence in Surgical Education, Research and Training (CESERT) from the Association for Surgical Education Foundation.


* Corresponding author. Tel.: +1-612-625-7911; fax: +1-612-626-0439.

E-mail address: chipm001@umn.edu

Manuscript received December 16, 2009; revised manuscript February 9, 2010

0002-9610/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.amjsurg.2010.02.006
Communicating with patients and their families is a basic skill required of all physicians and surgeons. Traditionally, these communication skills are developed by trainees “on the job” by emulating role models perceived to be effective, appropriate, and professional. Most surgeons experience little formal training and few opportunities for practice and feedback from either healthcare providers or family members. Communication about end-of-life (EOL) decisions and iatrogenic complications are two of the most challenging topics possible and are noted to be stressful for surgeons, as well as family members. Yet these conversations are common, particularly in intensive care units, where fewer than 5% of patients are able to talk about EOL decisions, and where half of all hospital deaths occur. It is not surprising, therefore, that professionalism, as well as interpersonal and communication skills, are 2 of the 6 competencies included by the Accreditation Council for Graduate Medical Education’s Outcome Project. Measuring these competencies, however, requires assessment tools, materials, and practices that generate reliable and valid scores.

We recently reported the creation of an Objective Structured Clinical Exam (OSCE) that uses standardized family members to simulate situations where surgeons discuss EOL issues and disclose an iatrogenic complication (disclosure of a complication [DOC]). This Family Conference OSCE allows residents to practice and develop communication skills and professional behaviors in a safe environment that allows for feedback and evaluation. Professional clinicians and standardized family members rated first-year (PG1) and third-year (PG3) surgical residents using a rating form that combined checklists and global evaluations.

As administered at the University of Minnesota in 2007, the Family Conference OSCE demonstrated high internal consistency reliability estimates, regardless of the case (EOL or DOC) or rater group. Interestingly, the family raters could discern no performance differences between the PG1s and PG3s on either case. Although PG3 residents assessed themselves significantly higher than the PG1s on EOL, and clinical raters’ scores trended upwards for PG3s for both cases, these data did not provide evidence of training year differences. We suspected small sample size and large, within-group variance may have limited our ability to detect true differences. However, we also began to question the assumption that resident training year could serve as a legitimate proxy for “expertise” in communicating bad news.

In an effort to increase our sample size and examine the reproducibility of our findings, we implemented the Family Conference OSCE at 5 other surgery residency training programs besides the University of Minnesota in 2008. Our primary goal was to determine the extent to which the OSCE generates scores that are reliable regardless of institution, type of rater (clinical vs standardized family member), and individual rater. Our secondary goal was to determine construct validity, that is, the extent to which scores measuring “professional communications” varied significantly by resident training year. We wished to further test the assumption that PG3 residents would, as a group, score higher than PG1 residents by virtue of having spent more time on critical care rotations and having more experience observing and participating in family conferences.

### Methods

#### Sites and participants

Based on the 2007 results of the Family Conference OSCE at the University of Minnesota, we calculated that we would need a sample of 60 residents (30 PG1, 30 PG3) to detect a difference in mean global evaluation scores of .74 (5-point scale) with 80% power. To enroll a sufficient number of residents, we recruited 5 additional institutions based on their interest and ability to perform the OSCE (see Table 1). Participating programs provided site directors, standardized family members, simulation facilities and staff, and clinical raters (ie, surgeons and nurses) to implement the OSCE. The institutional review boards of each site approved the study. The study was funded by a $28,508 grant from the Center for Excellence in Surgical Education, Research and Training (CESERT) from the Association for Surgical Education Foundation.

#### Cases

Cases were based on patient encounters from the surgical intensive care unit setting. One involved a terminal diagnosis and required the resident to present and discuss end-of-life issues with the patient’s husband and sister. The second required the resident to disclose a complication that significantly impeded a patient’s improvement to the primary care giver (sister) and daughter. Each case ran for 20 minutes. One month before the OSCE took place, a lecture on the topic of effective communication skills was presented at all

<table>
<thead>
<tr>
<th>Participating site</th>
<th>PG1</th>
<th>PG3</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Minnesota, Minneapolis, MN</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Hennepin County Medical Center, Minneapolis, MN*</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Mayo Clinic, Rochester, MN</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>University of Arizona, Tucson, AZ</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Medical College of Wisconsin, Milwaukee, WI</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Scott and White Clinic, Texas A&amp;M, Temple, TX</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>27</td>
</tr>
</tbody>
</table>

*Hennepin County Medical Center residents were tested with University of Minnesota residents.

Table 1 Participating general surgical residency programs and number of residents

![Table 1](image-url)
sites in a didactic setting. We made no attempt to characterize the residents’ prior experience or education with EOL or DOC discussions. Residents were provided with the details of the OSCE cases (ie, medical history and hospital course), as well as the task requirements, the night before the examination and also 5 minutes before the OSCE began. No background reading was provided, expected, or required. After each case, residents had 10 minutes to complete a self-assessment and OSCE evaluation form. All residents participated in post-exam debriefing sessions that were individually designed and conducted by participating sites.

Training for program sites

Site directors from all participating programs attended a day-long training conference in Minneapolis, MN. We discussed the background and rationale for the examination, we presented findings from our previous OSCE administrations, and we reviewed all the family and resident materials and rating instruments. We reviewed videotapes of previous OSCE performances and discussed the logistics of test administration and data collection. Additionally, we created a generic PowerPoint presentation about EOL and DOC issues that program leaders at each site used for the lecture they gave to residents in the month preceding the OSCE.

One or 2 of the authors (J.G.C. and/or C.C.S.) then traveled to each participating site just before the OSCE as local rater training took place. We did so to ensure consistency in family actors’ understandings and enactment of each case, and also in family and clinical raters’ scoring. Based on our previous experiences with the OSCE, family member actors were trained with detailed case scripts.4,5 Case scenarios were role played with 1 of the authors as the resident to allow the actors to experience the case.

Raters, rating tools, and data collection

Each case was recorded on video and rated by 2 clinical raters (a critical care–trained surgeon and an intensive care unit–experienced nurse) and the 2 family actors. Clinical raters evaluated the residents in real time and occasionally by video. The family actors rated the residents immediately following the encounter. Residents also rated their own performance after each case ended. Residents were evaluated via 2 rating tools (1 for EOL, 1 for DOC) that had been previously used and reported by the University of Minnesota.4,5 Scores from the rating tools were not combined, but analyzed separately, because they measure somewhat different communication skills. Additionally, case-by-trainee interactions have been fairly consistently reported in the literature.6,7 Each tool contains content-specific items with checklists and rating scales and several global evaluation items. Because the EOL and DOC rating tools are different lengths, to simplify evaluation and keep the results on the same metric, we used percent of possible points earned (PPPE) for analysis. Forms with more than 1 item missing were eliminated (1 from DOC, 2 from EOL). Residents from the University of Minnesota and Hennepin County Medical Center were tested together at the University campus; for these residents, scores were entered on an electronic system (B-Line). Evaluations conducted on all other sites were recorded on paper rating forms, which were collected by the participating sites. De-identified copies were sent to the University of Minnesota for data entry and data analysis.

Statistical analysis

We investigated the scale reliability of each rating tool by calculating Cronbach’s alpha for each case within each program site (see Table 2). Additionally, we examined the generalizability of scores across institutions, rater types (clinical or family), individual raters, and interactions with raters, by performing nested-design generalizability analyses for each case separately. We used generalizability theory to account for these potentially stacking, combining, and magnifying error possibilities. The benefit of generalizability theory is that it provides an “all-at-once way of revealing and comparing sources of error in a common metric.”8,9 In contrast, classical test theory considers only 1 type of potential measurement error at a time.

We calculated both the relative G-coefficient (suitable for a norm-referenced applications where the score of an individual is compared with his/her peers) and the absolute G-coefficient (suitable for criterion-referenced applications where the score of an individual is compared with some agreed upon absolute standard). As summarized by Downing, most educators seek a reliability coefficient of at least .90 for very high stake assessments, such as licensure or certification examinations.9 For more moderate stake assessments (eg, end-of-course or end-of-year summative ex-

### Table 2 Internal consistency reliability coefficients

<table>
<thead>
<tr>
<th>Institution</th>
<th>Cronbach’s alpha by case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DOC (n = 14 items)</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>12 .936</td>
</tr>
<tr>
<td>Hennepin County Medical Center</td>
<td>7 .867</td>
</tr>
<tr>
<td>University of Arizona</td>
<td>11 .913</td>
</tr>
<tr>
<td>Mayo Clinic Medical College of</td>
<td>13 .934</td>
</tr>
<tr>
<td>Wisconsin</td>
<td></td>
</tr>
<tr>
<td>Scott and White Clinic, Texas A&amp;M</td>
<td>10 .940</td>
</tr>
<tr>
<td></td>
<td>8 .924</td>
</tr>
</tbody>
</table>

DOC = disclosure of a complication scenario; EOL = end-of-life scenario.
aminations), reliabilities in the range of .80 to .89 are acceptable. For assessments with lower consequences (e.g., classroom-type assessments created and administered by local faculty), reliabilities in the range of .70 to .79 are acceptable.

Finally, we examined the construct validity of each case using repeated measures multivariate analysis of variance (MANOVA) based on the PG1 and PG3 scores. We did this for EOL and DOC separately.

Results

The Family Conference OSCE was administered to 61 residents (34 PG1s, 27 PG3s) at 6 surgery residency programs during the period from November 2007 through June 2008.

All categorical PG1s and all PG3s were scheduled for the examination.

Reliability/generalizability

High internal consistency reliability estimates were achieved at each site as shown by the Cronbach’s alpha coefficients in Table 2, which were greater than .9 for both cases at all but 1 site. As shown on Table 3, the overall relative generalizability coefficient for the EOL case was .890 and the absolute generalizability coefficient was .557. Individual differences attributable to residents accounted for greater than 55% of the variance, while individual raters and interactions with raters accounted for 36.1% of the variance. Interactions with rater type accounted for 6.9% of the variance. The remaining 1.2% of variance was accounted for by an institution-by-resident year interaction. For the DOC case, the overall relative generalizability coefficient was .716 and the absolute generalizability coefficient was .498. Differences attributable to individual residents accounted for nearly 50% of the variance in DOC scores. Interactions with different rater types accounted for 31.7% of the variance and individual raters and interactions with raters accounted for the remaining 18.5% of the variance.

Construct validity

As shown in Figure 1, the MANOVA analyses performed on EOL and DOC scores detected no significant difference between PG1 and PG3 residents for either case. Additionally, no significant interactions were found between rater types (clinical vs family) and training year. For DOC, the average score given by family members to PG1s was 76% versus 77% for PG3s; the average score given by clinical raters to PG1s was 74% versus 77% for PG3s. For EOL, family members’ scores were 71% (PG1s) and 73% (PG3s); clinical raters’ scores were 63% (PG1s) and 68% (PG3s). In other words, PG3s were not perceived to be more proficient than PG1s, no matter who was rating them.

Comments

There is a substantial and growing literature describing efforts to teach and assess communication skills to medical students and residents. However, most of the published articles describing OSCEs dealing specifically with the communication of “bad news” have involved physician

<table>
<thead>
<tr>
<th>Case</th>
<th>Relative G coefficient</th>
<th>Absolute G coefficient</th>
<th>Attributable variance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual resident</td>
<td>Rater type</td>
<td>Raters and interaction</td>
</tr>
<tr>
<td>EOL* (n = 61)</td>
<td>.890</td>
<td>.557</td>
<td>55.7%</td>
</tr>
<tr>
<td>DOC (n = 61)</td>
<td>.716</td>
<td>.498</td>
<td>49.8%</td>
</tr>
</tbody>
</table>

Table 3 Generalizability coefficients and attributable variance

Rater type: clinical or family rater, and interactions with rater type raters; and interaction: individual raters and the interaction with individual residents.

EOL = end-of-life scenario; DOC = disclosure of a complication scenario.

*1.2% of EOL variance was attributed to an institution-by-resident year interaction.
trained residents talking with patients; few have focused on communications with family members. The only known OSCE for surgical residents involves physician–patient communications in the clinic setting.12,13

Many training programs use faculty, peer, or nurse evaluations of residents’ interpersonal and communication skills to meet the Accreditation Council for Graduate Medical Education program requirements and their own local evaluation needs. The advantage of using an OSCE is that it standardizes the performance task and environment and theoretically allows for more objective ratings. Yet developing OSCE materials, rating tools, training and administration practices that support reliable and valid scores of communication skill remains a typically lengthy and iterative process.14

This study built on previous work by the University of Minnesota. Its goal was to test the feasibility of replicating the Family Conference OSCE at other training sites. Through a multi-institutional study, we hoped to overcome the limitations of small sample size to examine the generalizability of scores across settings, as well as rater types, and to better test the construct validity of the examination scores. One of the primary limitations of the current study is that it was not possible to blind all clinical raters to the training level of the residents. Therefore, we feel the scores given by family raters (who were blinded) are especially important.

In terms of generalizability, if one accepts relative versus absolute generalizability, the data suggest that both EOL and DOC examination scores were highly reliable and well within the parameters of use for moderate stakes assessment. In other words, residents who scored high in Minnesota would likely score similarly in Wisconsin or Arizona. However, the absolute generalizability scores were below the level requisite for summative examinations that use a set cut-off point. This is because a resident earning a certain percentage of possible points total in Minnesota could receive a somewhat different number (higher or lower) if he or she took the examination in Wisconsin. Nevertheless, the results suggest that this OSCE can be an appropriate and effective formative assessment for training purposes.

In terms of validating an underlying construct (“professional communication”), we were unable to find any evidence that either the EOL or DOC examination differentiated (what we assumed to be) novices (PG1s) from the more expert resident (PG3s). We do not feel this is due to a lack of statistical power, or to an examination that does not measure performances that are “true” (ie, what we intend to measure). Rather, we feel the assumption that PG3s are automatically better than PG1s by virtue of their training year is clearly open to challenge. Large standard deviations around mean scores for PG1s and PG3s, regardless of rater type, suggest strong individual differences are at work. Without evidence that interpersonal and communication skills and professionalism are systematically taught each year, and without evidence that residents get an increasing number of opportunities each year of training to observe and exercise leadership during family conferences, we should not assume PG3 residents are, as a group, more competent than PG1 residents. This conclusion therefore challenges us, and other investigators, to look for other ways to investigate the validity of our assessment processes. One method to explore the lack of construct validity would be to widen the difference between novices and experts. For example, administer the OSCE to early-year residents and attending physicians well versed in these types of communications and determine if the OSCE detects a difference.

A challenge for simulation in general and the use of standardized patients in particular is the expense of resources and personnel, which may be prohibitive to smaller training programs. Unanswered is if these communication skills could be practiced and evaluated with informal and inexpensive role-plays. Given the cost and time involved with developing OSCEs, the Medical College of Wisconsin decided to incorporate the EOL part of the OSCE into their PG2 palliative care curriculum as a formative assessment tool, and they also use aggregate data as a means for evaluating the curriculum. Additionally, they purposefully increased their junior residents’ exposure to EOL discussions with families in the intensive care unit.

Given their positive experience with the study, the University of Arizona elected to continue similar training sessions for all resident training levels, as well as their third-year medical students on their surgical clerkship. Educators at the Mayo Clinic planned to allow senior residents to lead family conferences in the intensive care unit and to use the Family Conference OSCE rating forms (self-assessment, plus attending feedback) to evaluate them. The annual partnership between the University of Minnesota and Hennepin County Medical Center surgical residencies was continued (for the third year) for the assessment of PG1s, PG3s, and critical care fellows; additionally, cardiothoracic surgery fellows from the University of Minnesota now take the examination.

While overall resident feedback has been positive, a common comment heard by at least some residents at all the sites (and as reported most clearly by the Texas A&M residents) was that the EOL scenario seemed “unrealistic” and difficult to completely engage. They felt a longer scenario that incorporated several stations (eg, perform a procedure, encounter a complication, move to another room and explain the situation to a family member) would make
the examination more realistic. We initially implemented such a progressive scenario in 2005 but found that while longer scenarios made it more realistic for the examinees, their skill levels were demonstrated within the first 20-minute station. Additional time and stations did little to change overall rater perception. As a practical, cost, and time-saving measure for both residents and raters, the lengthy, multistation scenarios were abandoned.

At the University of Minnesota, we hear this type of comment (“unrealistic”) most often during the debriefings that immediately follow the examination when residents are most anxious about their performance. Yet in year-end evaluation reviews in which residents present their portfolios of accomplishment, the Family Conference OSCE is nearly always mentioned as a seminal experience.

Conclusions

In conclusion, the Family Conference OSCE represents a feasible approach for assessing surgical trainees’ ability to deliver 2 types of bad news to family members in critical care settings. Scores are likely to be reliable if the standardized family members and clinical raters are well trained, if the lay actors are experienced in complex, semi-structured role plays, and if there are simulation staff (or others) who can provide needed administrative support. Scores from this OSCE are best used in formative assessment and teaching situations, and should not be used in isolation as summative assessment of ability. More research is needed to understand what the scores actually mean (construct validity), and their relationship to resident performance as demonstrated in actual family encounters.

Acknowledgment

The authors wish to express gratitude to Dr Jane Miller, Director, and Ann Wohl, Associate Director, and the lay actors and staff of the University of Minnesota Inter-professional Education Resource Center.

References

A Multi-Institutional Study of the Family Conference
Objective Structured Clinical Exam: A Reliable
Assessment of Professional Communication
Jeffrey G. Chipman, Travis P. Webb,
Mohsen Shabahang, Stephanie F. Heller,
Joan M. vanCamp, Amy L. Waer, Michael G. Luxenberg,
Matthew Christenson, and Connie C. Schmitz

A multi-institutional study provides support for the reliability of an Objective Structured Clinical Exam (OSCE) focusing on end-of-life decision-making and the disclosure of iatrogenic complications during simulated family conferences.