Surgical Education

How prepared are your interns to take calls?
Results of a multi-institutional study of simulated pages to prepare medical students for surgery internship

Adam C. Frischknecht, M.S.\textsuperscript{a,b}, Margaret L. Boehler, R.N., M.S.\textsuperscript{c,*}, Cathy J. Schwind, R.N., M.S.\textsuperscript{c}, Melissa E. Brunsvold, M.D.\textsuperscript{d,e}, Larry D. Gruppen, Ph.D.\textsuperscript{f}, Michael J. Brenner, M.D.\textsuperscript{g}, Linnea S. Hauge, Ph.D.\textsuperscript{b,f}

\textsuperscript{a}Michigan State University, College of Human Medicine, East Lansing, MI, USA; \textsuperscript{b}Department of Surgery, University of Michigan Medical School, Ann Arbor, MI, USA; \textsuperscript{c}Department of Surgery, Southern Illinois University School of Medicine, Springfield, IL, USA; \textsuperscript{d}Department of Surgery, University of Michigan, Ann Arbor, MI, USA; \textsuperscript{e}Division of Critical Care/Acute Care Surgery, Department of Surgery, University of Minnesota Medical School, Minneapolis, MN, USA; \textsuperscript{f}Department of Medical Education, \textsuperscript{g}Department of Otolaryngology – Head & Neck Surgery, University of Michigan Medical School, Ann Arbor, MI, USA

Abstract

BACKGROUND: This study evaluated a simulated pages curriculum that was developed to assess communication and clinical decision making in medical students and interns.

METHODS: A curriculum consisting of 14 simulated pages was administered across 5 institutions to 150 senior medical students. A 3-case subset was administered to interns who did not participate in the curriculum. Six expert surgeons identified critical fails and set passing scores for case-specific assessments using the Graphical Hofstee Method.

RESULTS: Participants in the curriculum demonstrated superior clinical decision making compared with non-participants across all cases scenarios ($P < .01$). Average medical student scores for clinical decision making were 46.9%. Global ratings averaged 6.0 for communication and 5.2 for patient care. Passing rates averaged 46%.

CONCLUSIONS: Participation in a mock page curriculum improved performance. The performance of participants based on expert standards set for simulated page performance highlight the need for innovative approaches to improve interns’ preparedness to take calls.

One of the critical challenges of a surgical intern’s 1st days on the wards is responding to pages about patients in need of immediate attention while on call. Despite the anxiety associated with this universal responsibility and the risks incurred with poor performance, little attention has been paid to preparing trainees to perform this task safely and effectively.\textsuperscript{1} The ability to function effectively as a healthcare
A team member is now regarded as a prerequisite skill for all medical professionals. Interprofessional communication is a critical aspect of safe team function, but general assessments of communication do not readily capture the nature of specific team interactions.

Physician–nurse telephone communication differs from other physician communication tasks in that the communication is not face-to-face, tends to be brief, and may involve patients the physician has not seen before. Yet, such consultations lead to critical decisions that affect patients’ care. Several studies cite the importance of effective physician–nurse communication in improving quality of care, patient outcomes, patient satisfaction, and in decreasing cost of care.2–5

Formal evaluation of physicians’ telephone communication skills has primarily targeted physician–patient telephone consultations,6,7 although some work has been done to identify key qualities or attributes of good physician–nurse phone communication.8 However, efforts to measure and improve the physician–nurse telephone communication skills have tended to focus on the nursing side of the interaction.9 One effort was aimed directly at improving medical students’ phone consultation skills and showed that exposure to simulated pages contributed to medical students’ clinical expertise and prepared them for internship.10,11

This study was designed to measure interprofessional communication and clinical decision making in phone consultations with nurses about common clinical challenges on the surgical unit. Instruments developed as part of this study were intended for competency assessment and standard setting in accordance with the recommendations of the Accreditation Council for Graduate Medical Education.12,13

Methods

Participants

This multi-institutional study received approval from institutional review boards at each participating institution and all participants provided informed consent. Senior medical students from 5 Liaison Committee on Medical Education accredited institutions participated in a formal mock page curriculum. This curriculum included 14 cases administered over the phone and was integrated throughout the 4-week surgical intern preparedness elective10,14–16 at each institution. Data for this study were collected during the administration of the mock page curriculum during the 2009 to 2010 and 2010 to 2011 academic years (Fig. 1). As a comparison group to the students in the formal curricula, we used a 3-case subset of the 14 mock page assessment cases to assess performance of 27 incoming interns who did not participate in the formal curriculum and/or any other type of intern preparedness curriculum. These incoming interns were assessed before they assumed any clinical responsibilities.

Case development

The development and implementation of the mock page curriculum is outlined in Fig. 1. A team of surgeons, residents, nurses, and medical education experts developed 14 cases (mock pages) representative of calls surgical interns might respond to on any given day. Cases targeted a specific clinical challenge and many of the cases included a specific communication challenge, for example, dealing with an inexperienced or confrontational nurse. Twelve cases involved physician–nurse communication and 2 cases included physician–patient communication. Sequencing of cases was designed to begin with less clinically challenging cases and the timing of each case throughout the 4 weeks was standardized for all sites.

Each case included an introductory script, standardized prompts, and pertinent information to be communicated to the student when queried. When a learner returned a page, the nurse or patient explained the reason for the call and then provided additional information as queried. Certain prompts were also scripted to ensure standardization of the case. Each case included 3 assessment components (Fig. 2):

1. Case-specific clinical decision-making checklist (scored to reflect whether the learner should or should not do each item): the total score on this component of the
assessment was reported as the percentage of positive points possible (ie, negative points were subtracted as demerits rather than being included in the calculation of total possible points).

2. Semantic differential scales for 8 constructs (modified from Reisdorff EJ, Hughes MJ, Castaneda, et al, 2006) related to communication: each construct was rated on a 7-point scale with higher ratings indicating greater levels of the constructs.

3. Global rating scales were used for rating the overall effectiveness of communication and the overall effectiveness of managing patient care on a scale of 0 (highly ineffective) to 10 (highly effective).

An external advisory group consisting of 3 surgery residency program directors from nonparticipating institutions reviewed the script, prompts, clinical information, and case-specific clinical decision-making assessment grid for each case. These experts provided comments and suggested edits to the cases. They also assigned scores to each case of the following constructs: (1) representativeness – a score of 1 was given for a case believed not to be representative of the clinical content that a new intern would likely encounter on call and a score of 5 was given for cases that were believed to be very representative of cases new interns would encounter; (2) complexity – simple cases were assigned a score of 1 and a score of 5 was given if the case was seen as being one of the most complex cases interns could encounter; (3) generalizability – a case that was viewed as being very specific to a single institution’s norms and practices was given a score of 1 and a case that was believed to be generalizable to most institutions was given a score of 5.

### Standardized paging team

A team of 7 nurses and 2 standardized patients administered the cases for all the pages completed in this study. Training for the nurses and standardized patients consisted of at least 3 practice runs of each assigned case. Learners for these practice sessions were recruited from the class of surgical interns at participating institutions and from a pool of 4th-year medical students at nonparticipating institutions. The nurse or standardized patient would give the case over the phone, while 3 to 4 members of the development team listened. At the conclusion of the case, everyone scored the case independently and then held consensus discussions about each assessment element. The nurse/standardized patient then practiced providing brief feedback to the learner. Revisions were made to the scenarios and the evaluation grids based on issues that arose during these practice sessions.

### Administration of the curriculum

At the beginning of the 4-week mock page program, learners participated in a 90-minute, interactive, case-based “common floor calls” training session before receiving any mock pages. The curriculum for this session was standardized across all participating institutions and covered all topics that were included in the mock page scenarios. The goals of the session were as follows:

1. to introduce learners to scenarios and procedures that they will likely encounter in their 1st nights on call as interns,
2. to describe effective and appropriate communication strategies for receiving and responding to pages from nurses, and
3. to reduce anxiety by rehearsing scenarios and reviewing relevant support mechanisms.

After the training session, the nurses and standardized patients implemented the 14 mock pages over the remainder
of the 4-week course. To maintain a realistic learning environment comparable with what the learners would experience as interns, learners were paged at unannounced times during the call schedule (8 am – 8 pm or 1 am – 5 am for night pages). When a learner returned a page, the nurse or standardized patient would begin the scenario script explaining the clinical situation. The learner was expected to assess the situation (query necessary clinical assessment information, laboratory results, vital signs, etc) and provide instructions for next steps (order labs or tests, inform the nurse they will come see the patient, and/or contact their senior). The case ended when the learner hung up the phone.

After the case ended, the nurse or standardized patient completed the assessment instruments. Then they called the learner back to provide brief (2 to 3 minutes), individual feedback about both the clinical management and the communication during the case. In addition, the learners participated in a weekly surgeon-led debriefing as a group at their respective institutions. In this debriefing, they reviewed and discussed proper management of all cases from that week.

**Surveys**

Mock page participants completed identical surveys before and after the course. These surveys queried the learners regarding their confidence and anxiety about responding to a nurse’s page on various common post-operative complications. Participants then rated their confidence on a scale from 0 to 10 (0 = not confident, 5 = somewhat confident, 10 = completely confident). They also rated their anxiety on a scale from 0 to 10 (0 = no anxiety at all, 5 = somewhat anxious, 10 = extremely anxious).

**Inter-rater reliability**

Inter-rater reliability was also studied for each of the 3 assessment components across each case. Twenty-one learners who completed all 14 cases were randomly selected from the 2010 cohort; audio recordings of their performances were used for inter-rater reliability studies. A 2nd rater (nurse or standardized patient) listened to the audio recordings of these learners in a given case. After listening to a recording, they completed the assessment instruments (case-specific clinical decision-making grid, semantic differentiation scale [SDS], and global ratings). Kappa was used to measure inter-rater reliability between the original and the 2nd raters for the clinical decision-component of the mock page program. They were oriented to the curriculum and the specific cases and then received instructions on the Graphical Hofstee Method of standard setting. After reviewing the cases, the group felt strongly that standard setting needed to take place in 2 phases: (1) identifying items on the grid that constitute “critical fails” and (2) setting a passing standard for the remaining items using the Graphical Hofstee Method. In both phases, judges were provided a summary of performance data from the 150 participants to calibrate their expectations and standards with actual performance data.

The judges were instructed to consider the following standard when making their judgments: a learner who passes a particular case would be allowed to assess the condition and initiate care with indirect supervision.

**Phase 1: critical fails**

The judges used the clinical decision-making grids from each case to independently identify items for each case that they considered to be “critical fails” if not done (or if done for the negative items). Results were tabulated and final “critical fails” were set if 4 of the 6 (67%) surgeons identified that item as a “critical fail.”

**Phase 2: pass/fail scores**

After “critical fails” were identified, individual learner scores were calculated for each clinical decision-making grid after excluding the critical fail items. In other words, assuming the learner avoided all “critical fail” items, where should the passing score be set for the remaining items. The judges were provided with a worksheet to use in making the 4 judgments that comprise the Graphical Hofstee Method for setting a passing score for each case. This worksheet informed the judges which items were identified as “critical fails,” the percentage of learners who would have failed based on those items alone, and the cumulative frequency distribution (CFD) plots of scores based on the remaining “noncritical fails.” The judges then made the 4 required judgments on the CFD plots as specified by the Graphical Hofstee Method.

Each of the 4 judgments was averaged across all judges. The passing standard for the noncritical fail portion of the grid was identified by finding the intersection between the CFD line and the line passing through the rectangle enclosed by the averaged judgments.

**Data analysis**

Data were collected through Microsoft Excel (Microsoft, Redmond, WA) and Qualtrics (Qualtrics, Provo, UT). Data were analyzed in PASW Statistics version 18 (IBM Corp, Armonk, NY). We calculated descriptive statistics for the assessment components from each case. Averages are reported as mean (standard deviation). Analysis of variance
<table>
<thead>
<tr>
<th>Clinical topic</th>
<th>Communication – semantic differentiation scale (1–7)†</th>
<th>Global ratings (0–10)‡</th>
<th>Management of patient care</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wound infection</td>
<td>Patient call</td>
<td>50 (17)</td>
<td>4.8 (1.2)</td>
</tr>
<tr>
<td>Abdominal distention</td>
<td>Inexperience</td>
<td>53 (20)</td>
<td>4.5 (1.2)</td>
</tr>
<tr>
<td>Dyspnea/pulmonary embolism</td>
<td>Unco-operative</td>
<td>44 (15)</td>
<td>2.7 (1.1)</td>
</tr>
<tr>
<td>Transfusion reaction</td>
<td>Inexperience</td>
<td>37 (16)</td>
<td>3.3 (1.2)</td>
</tr>
<tr>
<td>Patient pain</td>
<td>Worried</td>
<td>46 (15)</td>
<td>4.4 (1.0)</td>
</tr>
<tr>
<td>Wound hematoma</td>
<td>Unco-operative</td>
<td>26 (27)</td>
<td>3.3 (1.2)</td>
</tr>
<tr>
<td>Nausea/vomiting</td>
<td>Night page</td>
<td>53 (15)</td>
<td>4.8 (0.9)</td>
</tr>
<tr>
<td>Hypotension/hypovolemia</td>
<td>Night page</td>
<td>40 (14)</td>
<td>3.7 (1.3)</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td>Night page</td>
<td>54 (14)</td>
<td>5.1 (1.2)</td>
</tr>
<tr>
<td>Postoperative pain</td>
<td>Night page</td>
<td>43 (22)</td>
<td>4.2 (1.8)</td>
</tr>
<tr>
<td>Dehydration – ICU</td>
<td>Night page</td>
<td>53 (18)</td>
<td>5.3 (0.9)</td>
</tr>
<tr>
<td>Dyspnea/chest pain/MI</td>
<td>Family</td>
<td>54 (12)</td>
<td>4.8 (1.1)</td>
</tr>
<tr>
<td>Fever</td>
<td>Unco-operative</td>
<td>58 (17)</td>
<td>4.9 (1.7)</td>
</tr>
<tr>
<td>Acute alcohol withdrawal</td>
<td>Unco-operative</td>
<td>46 (15)</td>
<td>4.8 (1.2)</td>
</tr>
</tbody>
</table>

ICU = intensive care unit; MI = myocardial infarction; SD = standard deviation.
*Several cases included deliberate communication challenges in addition to clinical challenge.
†The semantic differentiation scale is a 7-point scale with higher ratings indicating greater levels of the particular construct.
‡Effectiveness of communication and clinical care was scored from 1 = highly ineffective to 10 = highly effective.
was applied to compare assessments and survey responses between the corresponding groups. Criterion statistical significance level was defined as $P$ value less than .05. Effect sizes (Cohen’s $d$) were calculated for differences between mock page participants and nonparticipants on the subset cases. Effect sizes were considered small ($d > .2$), moderate ($d > .5$), and large ($d > .8$). Paired t tests were used to compare mock page participants’ pre- and postcourse responses with the confidence and anxiety surveys.

Results

The external reviewers agreed that the cases were representative of the clinical content that a new intern will likely encounter on call (range: 3.3 to 5.0 out of 5) and generalizable across institutions (range: 4.0 to 5.0 out of 5). The reviewers also indicated that the cases represented a range of complexity levels (range: 1.7 to 5.0 out of 5). Cases were revised based on recommendations from these reviewers.

Inter-rater reliability was good for each of the evaluation grids with kappa averaging .83 and ranging from .75 for the wound hematoma case to .91 for the chest pain/MI and the pulmonary embolism cases. The communication scales (SDS and global communication ratings) were less reliable, highlighting the difficulty in consistently and accurately assessing communication skills. ICC for the SDS averaged .36 and ranged from .17 in the vomiting/nausea case to .56 for the dehydration – intensive care unit (ICU) case. The ICC for global communication ratings averaged .36, ranging from .07 in the patient wound infection case to .75 in the dehydration – ICU case. The global clinical management ICC averaged .64 and ranged from .13 in the acute alcohol withdrawal case to .91 in the fever case.

Mock page performance

One-hundred fifty medical students participated in the mock page curriculum during their 4th year in 2010 ($n = 79$) or 2011 ($n = 71$). These students were distributed across 5 institutions with a similar number of students at each site.

Table 1 presents the descriptive summaries for the individual components of the assessment instrument across all 14 cases. On average, 141 learners completed each case. Average scores for case-specific clinical decision making were 46.9% (8.6%). Average global ratings were 6.0 (1.2) for effectiveness of communication and 5.2 (1.2) for effectiveness of managing patient care. Subscale scores from the SDS averaged 4.4 (.7).

Twenty-seven nonparticipant incoming interns completed the transfusion reaction and pulmonary embolism – ICU cases in 2010 and 2011. Eight of these incoming interns also completed the patient pain case in 2010. The standardized patient was unavailable to administer the case to incoming interns in 2011, so this case was excluded that year rather than introducing a 2nd standardized patient as a confounding factor. Table 2 shows the performance of the nonparticipants compared with mock page participants on these 3 cases. The mock page participants scored significantly higher than the nonparticipants on all assessment measures for the pulmonary embolism case. They also scored significantly higher on some assessment measures for the transfusion reaction case (clinical decision making, precision SDS subscale, and directing SDS subscale) and the patient pain case (clinical decision making, instructive SDS subscale, and engaged SDS subscale). Effect sizes for significant differences were all moderate to large.

### Table 2 Summary of performance of mock page participants compared with non-participants

<table>
<thead>
<tr>
<th>Case description</th>
<th>Participants</th>
<th></th>
<th>Non-participants</th>
<th></th>
<th>$P$ value</th>
<th>Effect size (Cohen’s $d$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transfusion reaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical decision making</td>
<td>37.0%</td>
<td>16.0%</td>
<td>28.0%</td>
<td>13.2%</td>
<td>.01</td>
<td>.58</td>
</tr>
<tr>
<td>Semantic differentiation scale</td>
<td>3.8</td>
<td>.9</td>
<td>3.5</td>
<td>1.0</td>
<td>.16</td>
<td>.29</td>
</tr>
<tr>
<td>Global communication</td>
<td>4.2</td>
<td>1.6</td>
<td>4.1</td>
<td>1.6</td>
<td>.64</td>
<td>.10</td>
</tr>
<tr>
<td>Global patient care</td>
<td>3.4</td>
<td>1.8</td>
<td>2.7</td>
<td>1.8</td>
<td>.09</td>
<td>.36</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical decision making</td>
<td>53.9%</td>
<td>13.7%</td>
<td>30.9%</td>
<td>15.0%</td>
<td>&lt;.001</td>
<td>1.66</td>
</tr>
<tr>
<td>Semantic differentiation scale</td>
<td>5.2</td>
<td>.7</td>
<td>3.9</td>
<td>1.2</td>
<td>&lt;.001</td>
<td>1.55</td>
</tr>
<tr>
<td>Global communication</td>
<td>6.8</td>
<td>1.5</td>
<td>6.0</td>
<td>1.9</td>
<td>.03</td>
<td>.46</td>
</tr>
<tr>
<td>Global patient care</td>
<td>5.5</td>
<td>2.0</td>
<td>3.7</td>
<td>2.4</td>
<td>&lt;.001</td>
<td>.89</td>
</tr>
<tr>
<td>Patient pain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical decision making</td>
<td>45.6%</td>
<td>15.2%</td>
<td>28.9%</td>
<td>12.5%</td>
<td>.003</td>
<td>1.10</td>
</tr>
<tr>
<td>Semantic differentiation scale</td>
<td>4.2</td>
<td>.9</td>
<td>3.8</td>
<td>1.4</td>
<td>.14</td>
<td>.54</td>
</tr>
<tr>
<td>Global communication</td>
<td>5.3</td>
<td>1.7</td>
<td>5.3</td>
<td>1.8</td>
<td>.88</td>
<td>.06</td>
</tr>
<tr>
<td>Global patient care</td>
<td>5.3</td>
<td>1.7</td>
<td>4.5</td>
<td>2.3</td>
<td>.18</td>
<td>.49</td>
</tr>
</tbody>
</table>

SD = standard deviation.

Semantic differentiation scale included subscales from 1 to 7 for 8 different constructs of interpersonal and communication skills. Higher scores indicated a greater demonstration of the particular construct.
Survey results

Pre- and postcourse confidence and anxiety surveys were returned from 132 mock page participants (88%). Average confidence ratings improved significantly between the precourse 4.2 (1.6) and postcourse 7.4 (1.1) ($P < .000, d = 2.3$). Likewise, average anxiety ratings decreased significantly from precourse 6.1 (1.6) to postcourse 4.2 (1.8) ($P < .000, d = 1.1$). Average self-assessment of the 8 constructs in the SDS also improved significantly from precourse 4.8 (.8) to postcourse 5.8 (.6) ($P < .000, d = 1.4$).

Standard setting

Table 3 displays the defined standards and summarizes their retrospective application with the existing performance data. For each case, an average of 2.1 (1.5) items were identified as “critical fails” and would have resulted in average failure rates of 13% (10%) based on “critical fails” alone. Passing scores for the remaining items set for individual cases averaged 36% (7%), resulting in average total failure rates (critical fails and/or low scores) of 46% (14%).

Comments

This study demonstrates the implementation of a formal mock page curriculum integrated into existing intern preparedness courses across 5 institutions. Work relating to this curriculum has been adopted as part of the National Curriculum, a resident preparatory course developed in collaboration with the American College of Surgeons, the Association of Program Directors in Surgery, and the Association of Surgical Education now being piloted at 40 medical schools. Therefore, the validation and field work included in this analysis provide important data regarding the tool. Participation in this mock page curriculum improved performance in comparisons with non-participants. Previous single-institution studies have demonstrated the use of mock pages and reported similar increases in learners’ confidence. However, these previous studies focused solely on clinical management and did not include assessments of interprofessional communication. The addition of assessments of interprofessional communication provides a more complete picture of the physician–nurse and physician–patient telephone communication dynamic. Getting the diagnosis right or knowing which tests to order does not obviate the need to effectively communicate with members of the healthcare team who can initiate management plans and provide essential help in providing safe and effective patient care.

Making the task as realistic as possible not only facilitates transfer of learning from the simulated setting to real life, but also the use of assessments in the simulated setting to inform predictions of how learners will perform in actual clinical settings. By initiating the cases via pagers, the learners face the realistic requirement of responding to unannounced clinical pages that must be answered in a timely manner. Communication challenges added to several cases enhanced the realistic nature of the exercise and allowed for training and assessment of clinical decision-making skills and communication skills in less-than-ideal communication settings.

Average clinical decision-making scores were below 50%, suggesting that senior medical students transitioning...
to internship may not be prepared to respond to common pages from the surgical units. These results are disconcerting given the current practice in many institutions to assign interns to take call as soon as they become interns. New requirements for increased supervision\(^1\) as well as increasing requirements of objective measures of competence\(^1\) may contribute to improved intern performance and improved patient care, but 1st require the development and application of appropriate training and assessment tools. The mock page assessments presented here are a viable option for assessing intern performance in managing common clinical problems and identifying specific areas for remediation. Additionally, these tools provide a means to assess communication and interprofessional communication and emphasize the importance of developing strong interdisciplinary collaboration among members of the healthcare team. The ability to teach and assess trainees in several of the Accreditation Council for Graduate Medical Education Core Competencies simultaneously makes this mock page curriculum a valuable model for medical education and assessment. Patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice could all be taught and assessed within the context of the mock page curriculum.

Our most striking results were with the comparisons between mock page participants and nonparticipant interns on the subset of 3 cases. These comparisons suggested that the mock page curriculum did improve learner’s clinical decision making and effective communication in these cases. Another indication of the effectiveness of the curriculum is seen in the self-assessment measures of confidence and anxiety. Both measures improved significantly between pre- and postcourse queries (confidence increased, anxiety decreased). Because the responsibility of taking call is one of the stresses interns must manage as they assume their clinical responsibilities, the mock page experience provides an important opportunity for interns to practice this task in a safe, low-stakes environment.

With the combination of “critical fails” and the Hofstee cutpoint scores, the total percentage of learner failures ranged from 31% to 71%. This level of failure may be acceptable in a formative educational context in which there is opportunity to remediate learners who failed. In a summative context, in which failing learners would need to be sequestered from clinical responsibilities while they acquire the skills and experience to reach competence, these high rates of failure would indicate that there needs to be either a careful look at the judgment processes used by the judges or there needs to be significant investment in earlier curricular interventions to address this competence deficiency. Because cases with increasing complexity and communication challenges were deliberately scheduled at the end of the curriculum, the implications of high failure rates throughout the curriculum cannot be interpreted without further study.

A few limitations of this study should be considered. Given the multi-institutional nature of the study, specific teaching provided during the introductory common calls session and the weekly debriefings varied from one institution to another. Standardized instructions were given to all instructors to minimize this potential source of variation. However, this institutional variation can also be seen as a strength in providing more generalizable results.

Disagreement between raters for the communication scales and the clinical management scales may reflect differences in understanding of the specific constructs being assessed. Additional rater training or more regular checks of rater agreement may lead to improved reliability of these instruments.

Despite the efforts to make the experience as realistic as possible, this study still only captures learner’s performance in a simulated environment. As such, the ability to use these assessments to predict future performance during residency has not been validated. A pilot study was conducted in 2010 to collect in situ performance data from mock page participants and their peers during the 1st weeks of internship. However, several factors diminished the value of these early data. First, follow-up became increasingly difficult as the learners graduated and moved throughout the country to begin their internships. Logistically, only mock page participants who matched to 1 of our 5 participating institutions were available for inclusion in this in situ pilot study. Additionally, attempts were made to recruit nurses to serve as evaluators on all the surgical units where the interns would be assigned. However, nurses were not able to assess the interns when they did not have to page a particular intern on a given night. The number of nurses recruited and the distribution of their assignments across institutions and units limited nurse training on the assessment instruments. Small sample sizes (n = 6 mock page participants and n = 16 nonparticipant controls) resulted in an underpowered analysis of the impact of the mock page curriculum on actual clinical performance as interns.

In conclusion, participation in this mock page exercise improved performance on a simulated page intern task. The mock page exercise is thus an effective means of learning to appropriately respond to pages. However, a 2-phase standard-setting process resulted in standards for summative assessment of mock page performance that would still result in a substantial percentage of learners requiring remediation before being deemed ready to assess the given conditions and initiate care with indirect supervision. The standards include consideration of both “critical fails” as well as overall performance. Participation in this curriculum increased learner’s confidence and decreased anxiety in responding to pages from nurses on the surgical unit. This mock page curriculum has been adopted as a major portion of the national curriculum for medical students entering surgery residency developed in 2012 in collaboration with the American College of Surgeons, the Association of Surgical Education, and the Association of Program Directors in Surgery.
References