Closing the Gap in Operative Performance Between Novices and Experts: Does Harder Mean Better for Laparoscopic Simulator Training?

Dimitrios Stefanidis, MD, PhD, James R Korndorffer Jr, MD, FACS, Sarah Markley, BS, Rafael Sierra, MD, B Todd Heniford, MD, FACS, Daniel J Scott, MD, FACS

BACKGROUND: We have previously shown that reaching expert performance on an fundamentals of laparoscopic surgery (FLS)-type simulator model for laparoscopic suturing results in measurable improvement during an actual operation; trained novices, however, demonstrate inferior operative performance compared with experts. We hypothesized that simulator training under more difficult and realistic conditions would enhance the operative performance of novices.

STUDY DESIGN: Medical students (n = 32) participated in an IRB-approved, randomized, controlled trial. All participants were pretested in laparoscopic suturing on a previously validated porcine Nissen model and were randomized into three groups: group I (n = 6) received no training, group II (n = 13) trained on the FLS videotrainer model until a previously published proficiency score (512) was achieved on 2 consecutive and 10 additional attempts, group III (n = 13) trained to the same goal but had to practice in a constrained space, with a shorter suture, starting with a dropped needle, and listening to operating room noise. Training workload was measured with the validated NASA-TLX (Task Load Index) questionnaire after each training session. All groups were posttested on the porcine model. Results were compared using ANOVA; p < 0.05 was considered significant.

RESULTS: All group II and III participants reached the training goal. At posttesting, group II and group III participants performed similarly, but substantially better than group I did (210 ± 140 versus 218 ± 139 versus 0 ± 0, respectively; p < 0.001). Compared with group II, group III participants trained longer (329 ± 71 minutes versus 239 ± 69 minutes, p < 0.001), performed more repetitions (81 ± 15 versus 59 ± 14, p < 0.001), and their workload improved less by the end of training (5% versus 23%, p < 0.001).

CONCLUSIONS: Proficiency-based simulator training reliably results in improved operative performance. Although increasing the level of training difficulty increased trainees’ workload, the strategy we used in this study did not enhance their operative performance. Other methods for curriculum optimization are needed. (J Am Coll Surg 2007;205:307–313. © 2007 by the American College of Surgeons)

Simulation is gaining widespread acceptance for surgical training. It allows inexperienced trainees to acquire new skills through repetitive practice in a safe, nonthreatening environment before patient encounters in the operating room. Although multiple simulators have been developed and validated for training,1-3 simulator curricula need additional development. And even though efforts of surgical organizations to create a national curriculum are ongoing,4 no uniformly accepted curriculum is currently available that will provide trainees with optimal skill acquisition.

Proficiency-based simulator curricula have proved effective in improving the operative performance of trainees.2,5 Nevertheless, our group has clearly demonstrated that although individuals trained to proficiency on sim-
ulators uniformly outperform controls in the operating room, their performance falls well below that of experts in the real life environment. \(^5\) In a previously published randomized controlled trial, we demonstrated that nine residents who trained to proficiency in laparoscopic suturing using an fundamentals of laparoscopic surgery (FLS)-type videotrainer achieved a score of 389 ± 70 in a live porcine Nissen model, and their performance was better than that of residents who did not train on the simulator (217 ± 140; \(p < 0.001\)), but was still substantially inferior to that of laparoscopic experts (504 ± 9; \(p < 0.001\)). \(^5\)

The cause of this performance gap that appears to be unmasked under operating room conditions may be multifaceted. One of the main contributing factors might be that training on a low fidelity simulator may be too easy compared with the demanding operating room environment, which can be stressful, offers many distractions, is characterized by the element of the unexpected, and imposes motion artifact and spatial constraints. Indeed, studies have shown that the operating room represents a noisy environment that can be as loud as 80 to 85 dB\(^6\) and that cognitive distractions negatively influence the performance of laparoscopic tasks. \(^7\) In addition, it has been suggested that increasing training difficulty and variability improves skill acquisition. \(^8\)-\(^10\)

In an attempt to close this performance gap between simulator-trained novices and experts, we hypothesized that novices trained to proficiency under more difficult and variable simulator conditions would outperform conventionally trained novices in the operating room.

**METHODS**

Medical students (\(n = 32\)) voluntarily participated in an IRB-approved, randomized, controlled trial that was conducted at the Tulane Center for Minimally Invasive Surgery in New Orleans, LA. All participants completed a questionnaire about demographics and earlier laparoscopic and simulator experience and watched a video tutorial of laparoscopic suturing and knot tying. Participants were then asked to place one gastrogastric suture on a previously validated live porcine Nissen model, \(^3\) which served as our baseline test. Students were then stratified according to their performance and randomized into three groups: group I (\(n = 6\)) received no training and served as controls, and groups II (\(n = 13\)) and III (\(n = 13\)) trained to proficiency in laparoscopic suturing on an FLS-type videotrainer model, as previously described. \(^3\) Stratification was in blocks of five students who were randomized in a 4 (training) to 1 (control) fashion.

Group II trained until they achieved a previously published expert derived performance level (score of 512)\(^5\) on 2 consecutive plus 10 additional attempts. Group III trained until they achieved the same expert level (512) on two consecutive attempts under regular conditions, but then had to perform the task in a constrained space (box trainer space constricted by 50%), had to listen to operating room noise through headphones (the noise had been recorded in a busy operating room), had to practice with a shorter suture (4 inches long instead of the standard 6-inch suture used for conventional training), and had to start with a dropped needle, whose tip was facing away from the FLS model (normally participants would introduce the needle and suture with their instruments through the trocars and preorient the needle to face the target). These four difficult conditions were introduced gradually into the curriculum; participants had to achieve the expert level on two consecutive attempts with the constrained space and operating room noise first and then with the short suture and dropped needle start. To complete the curriculum, group III participants had to achieve the expert level on two consecutive attempts plus 10 additional times, practicing with all conditions. Trainees were made aware before training of the required curriculum and of the nature and purpose of the four additional training conditions.

After completion of training, all participants were brought back for a retention test on the simulator under regular conditions. All groups (including the control group) were again tested on the live porcine Nissen model to assess transfer of simulator-acquired skill to the operating room (Fig. 1).

All trainees received augmented feedback during each training session by the same instructor and had access to video tutorials to optimize skill acquisition. Training was in a distributed fashion, with sessions occurring once or twice a week; the duration of each session did not exceed 1 hour. Cut-off time for each repetition was 10 minutes. After each training and testing session, participants completed the NASA-TLX (Task Load Index) workload assessment questionnaire. \(^11\) At study completion, participants also completed a questionnaire about their training experience.
Performance during training and testing sessions was assessed with objective scores based on time and errors using the following previously published formula:

\[600 - (\text{time} + \text{accuracy error} \times 10 + \text{incomplete knot approximation error}) \times 10 + \text{security error} \times 10\]

For workload assessment, the validated NASA-TLX questionnaire was administered. This instrument uses a 20-point visual analogue scale to measure the mental, physical, and temporal demands of a task and the effort, frustration, and perceived performance of the subject performing the task.

**Statistical analysis**

Training and testing performance and workload data, baseline, and study completion questionnaire data were entered into a database and analyzed. Statistical analysis
was conducted using the Sigma Stat 3.0 statistical software (SPSS Inc); ANOVA and t-test were used to compare between-group differences and paired t-tests were used for within-group comparisons. A p value < 0.05 was considered statistically significant. The training group size (n = 13) was chosen to allow detection of a performance score difference of 80 between the training groups during the testing sessions with a power of 0.8 and an alpha level of 0.05 based on a performance standard deviation of 70. The standard deviation was based on previous results on the same model, and the chosen score difference corresponded to 15% of the expert performance level.

RESULTS
Participant age was 25 ± 3 years, 31% were women and 97% were right hand dominant. No participant had earlier experience with laparoscopy or simulators. On a 1-to-10 Likert scale, earlier participant exposure to video games or billiards was 5 (range 2 to 10) and was similar for all groups.

No participant was able to complete the suturing and knot tying task in the live porcine Nissen model at baseline within the allotted 10 minutes. One participant, who was randomized to group III, did not attend any training session and was excluded from the study. All participants who started training achieved proficiency and completed the study. Compared with group II participants, group III participants trained longer (329 ± 71 minutes versus 239 ± 69 minutes, p < 0.001), performed more repetitions (81 ± 15 versus 59 ± 14, p < 0.001), and necessitated more training sessions (5.1 ± 0.9 versus 4.2 ± 1.7, p < 0.05).

The NASA workload scores were similar between the groups at the beginning and at the end of training. Nevertheless, once the four conditions were introduced in the training of group III participants after they had demonstrated initial proficiency (achievement of proficiency level on two consecutive attempts), group III workload scores increased by 33% compared with those of group II (88 ± 8 versus 66 ± 12, respectively; p < 0.001, Fig. 2). In addition, compared with the beginning of training, participant workload decreased substantially at training completion for group II (83 ± 11 versus 64 ± 16; p < 0.01), but not for group III (76 ± 12 versus 72 ± 22; p = not significant [ns]). This decrease in workload scores was substantially higher for group II compared with group III participants (38% versus 5%, respectively; p < 0.001).

The retention test on the simulator and the posttraining test in the operating room were administered on the same day, with an average interval of 8.4 ± 7.7 days between testing and training completion. Compared with group III, group II participants required substantially longer to complete the task (100 ± 31 seconds versus 81 ± 21 seconds, p < 0.05) and tended to have lower performance scores (500 ± 31 versus 518 ± 22, p = 0.09) during the simulator retention test. In addition, fewer group II participants achieved the proficiency score on their first attempt compared with group III (31% versus 83%, p < 0.05).

In contrast, the performance of group II was similar to that of group III (210 ± 140 versus 218 ± 139, respectively; p = ns) in the live porcine model. Nonetheless, both training groups achieved notably better scores compared with the control group (p < 0.001, Fig. 3). During both the baseline and posttraining porcine tests, there were no differences in the NASA-TLX workload scores between the three groups. Nevertheless, workload scores showed a trend toward improvement for all groups at the posttraining test compared with the baseline test. This change was statistically significant only for the group II workload scores (93 ± 18 at baseline versus 75 ± 22 after training completion; p < 0.05) and not for group I (87 ± 11 versus 80 ± 22, p = ns) or group III (90 ± 18 versus 84 ± 22, p = ns).
Group III trainees required more repetitions to reach the proficiency level with the constraint space and noise than with short suture and dropped needle start, but this difference was not statistically significant (19/100 vs 13/100, respectively; $p = ns$). At the end of their training, group III trainees ranked the difficulty of the added conditions as shown in Table 1. On a 1-to-10 Likert scale, group III participants rated training with the added conditions to be 6.2 ± 1.7 times more difficult compared with regular training. Seven (58%) participants thought that their training under more difficult conditions had a positive effect on their learning, three (25%) perceived a negative effect, and two (17%) were uncertain. Five (42%) participants would have chosen all additional conditions for training, three (25%) would have chosen the constrained space, three (25%) the short suture, and one (8%) the dropped needle start; seven (58%) participants would have preferred to train under even more difficult conditions. On the other hand, five (38%) group II participants would have preferred to train under the more difficult conditions, two (15%) preferred not to have the additional conditions, and six (46%) were uncertain.

**DISCUSSION**

We undertook this study because our previous work demonstrated that although proficiency-based laparoscopic suturing training on simulators is effective, there still remains a substantial gap between trainees and experts in the operating room environment. This study demonstrated that increasing the difficulty level of simulator training led to improved trainee short-term skill retention on the simulator, but did not improve operating room performance when compared with performance of conventionally trained individuals. In addition, this occurred at the expense of longer training times and higher trainee workloads. In other words, novices who trained conventionally on the simulator achieved the same operating room performance as did those who trained harder and longer.

A recent systematic review about the effect of high-fidelity medical simulations on learning indicated that learning was enhanced when trainees had the opportunity to practice with progressively increasing levels of difficulty. Using the minimally invasive surgical trainer—virtual reality simulator, Ali and colleagues demonstrated that training on the medium level resulted in improved skill acquisition when compared with training on the easy level, and Aggarwal and associates suggested that progressively increasing levels of difficulty were optimal for curriculum design. Our study supported these findings; participants who trained under more difficult simulator conditions outperformed conventionally trained individuals on the simulator according to task completion time and ability to reach proficiency at first attempt.

Unlike the previously mentioned studies, however, this was the first study that examined the transferability of an increased difficulty laparoscopic simulator curriculum to the operating room environment. In contrast to the small performance improvements seen on the simulator after more difficult training, we did not demonstrate a benefit in the more realistic and demanding assessment environment of the operating room.

This lack of operative performance improvement despite the increased difficulty during training may have been related to simulator fidelity. It is possible that the learning benefit that our low fidelity videotrainer model provided trainees with had already been maximized by our rigorous standard proficiency-based curriculum. Our findings may also imply that small performance changes in a low fidelity simulator do not translate to the operating room. In addition, box trainers offer a much more controlled training environment compared with the operating room, and for this reason alone, may be less discriminatory of novices versus experts.
In an effort to increase the realism of the simulation, we chose four clinically relevant conditions (constrained space, operating room noise, short suture, and dropped needle start). Interestingly, constrained space and short suture were perceived as the most challenging conditions by participants, and operating room noise as the least. This may not come as a surprise because auditory stimuli compete for different attentional resources than those required for laparoscopic suturing and as such, are less likely to disrupt the performance of this task. By choosing these four conditions, our goal was not to increase the task difficulty (eg, by requiring greater precision or faster task completion), but rather to modulate training conditions using multiple disparate methods that better reproduce and more globally represent situations likely to be encountered by trainees in clinical practice. So our study cannot answer the question of whether targeted increases in task difficulty, as can be achieved with some virtual reality simulators, may augment learning. Nevertheless, we believe that incorporation of such conditions should offer additional experiences to trainees that improve their readiness for the operating room. Indeed, the improvement seen in trainee simulator performance and the favorable trainee ratings at least partially support their validity. In addition, the 33% increase of NASA-TLX workload scores after introduction of these conditions was indicative of a true increase of the training difficulty level. Even so, the additional features of this simulation failed to yield a measurable objective difference in the operating room. Although incorporating additional conditions or using a higher fidelity model, such as ex vivo porcine specimens, might have had a larger impact, other studies have implied that skill acquisition is not influenced by simulator fidelity. Grober and co-workers examined the impact of bench model fidelity on the microvascular skills of 50 junior residents and concluded that skills training on low fidelity simulators was equally effective as training on high fidelity simulators.

On the other hand, the increased difficulty model may have proved more beneficial with additional training. Even though training duration was 37% longer for group III, these participants may not have fully mastered suturing under the difficult conditions according to the NASA-TLX workload scores. Although group II demonstrated a marked 23% decrease in workload scores between the first and last training sessions, group III achieved only a nonsignificant 5% fall, in part because of a higher workload imposed by the more difficult training conditions (Fig. 2). Because workload reduction during training reflects increasing comfort with a task and, consequently, its mastery, the lack of marked improvement of group III NASA-TLX scores by the end of training may be reflective of the need for additional training. This begs the question whether the traditional performance metrics of time and errors that we used in this study are sensitive enough to truly define proficiency; additional metrics like the NASA-TLX workload questionnaire may be a useful adjunct.

Performance anxiety may have been an additional factor because both groups were familiar with the study design and hypothesis. Participants who trained under the more difficult conditions may have felt pressure to outperform their counterparts in the operating room, which, in turn, might have led to increased anxiety and suboptimal performance. Although we did not directly assess anxiety in the operating room, the NASA-TLX workload scores may provide indirect evidence in support of this argument. Although group II workload scores improved considerably between the porcine pre- and post-tests, the same was not true for group III scores. So, group III members had to work harder during the posttest (a factor possibly related to anxiety), and their performance may have suffered. Indeed, a study by Wetzel and co-authors demonstrated that stress can impair surgeon judgment, decision making, and communication—all factors that may negatively affect performance.

### Table 1. Ranks of the Difficulty of Added Conditions

<table>
<thead>
<tr>
<th>Difficulty rank</th>
<th>n</th>
<th>%</th>
<th>n</th>
<th>%</th>
<th>n</th>
<th>%</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constrained space</td>
<td>5</td>
<td>42</td>
<td>3</td>
<td>25</td>
<td>4</td>
<td>33</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Short suture</td>
<td>4</td>
<td>33</td>
<td>3</td>
<td>25</td>
<td>4</td>
<td>33</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Dropped needle</td>
<td>2</td>
<td>17</td>
<td>6</td>
<td>50</td>
<td>3</td>
<td>25</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Operating room noise</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td></td>
<td>2</td>
<td>17</td>
<td>9</td>
<td>75</td>
</tr>
</tbody>
</table>

1, most difficult; 4, easiest.
Despite the shortcomings of this study, important data were established as we strive to make additional improvements in simulation. Our results are encouraging in supporting the benefit of more rigorous training, as demonstrated by improvements in simulator performance and positive trainee perceptions. Additionally, although differences between groups were not detected, this study confirmed earlier studies by clearly demonstrating the transfer of skills from the bench to the operating room after simulator training. Additional work is still needed to better define metrics and simulator design in terms of optimizing proficiency-based curricula for advanced skills such as laparoscopic suturing. Additional investigations are warranted in hopes of realizing a more profound impact on skill acquisition using simulators before clinical experiences. Our ultimate goal is to allow trainees to reach performance levels similar to those of experts in the demanding milieu of actual operating room environments.

Author Contributions

Study conception and design: Stefanidis, Korndorffer, Scott
Acquisition of data: Stefanidis, Markley, Sierra
Analysis and interpretation of data: Stefanidis, Heniford, Scott
Drafting of manuscript: Stefanidis, Markley, Sierra
Critical revision: Stefanidis, Korndorffer, Heniford, Scott

REFERENCES