Simulator Training to Automaticity Leads to Improved Skill Transfer Compared With Traditional Proficiency-Based Training

A Randomized Controlled Trial

Dimitrios Stefanidis, MD, PhD; Mark W. Scerbo, PhD; Paul N. Montero, MD; Christina E. Acker, BA; Warren D. Smith, PhD


Abstract and Introduction

Abstract

Objective: We hypothesized that novices will perform better in the operating room after simulator training to automaticity compared with traditional proficiency based training (current standard training paradigm).

Background: Simulator-acquired skill translates to the operating room, but the skill transfer is incomplete. Secondary task metrics reflect the ability of trainees to multitask (automaticity) and may improve performance assessment on simulators and skill transfer by indicating when learning is complete.

Methods: Novices (N = 30) were enrolled in an IRB-approved, blinded, randomized, controlled trial. Participants were randomized into an intervention (n = 20) and a control (n = 10) group. The intervention group practiced on the FLS suturing task until they achieved expert levels of time and errors (proficiency), were tested on a live porcine fundoplication model, continued simulator training until they achieved expert levels on a visual spatial secondary task (automaticity) and were retested on the operating room (OR) model. The control group participated only during testing sessions. Performance scores were compared within and between groups during testing sessions.

Results: Intervention group participants achieved proficiency after 54 ± 14 and automaticity after additional 109 ± 57 repetitions. Participants achieved better scores in the OR after automaticity training [345 (range, 0–537)] compared with after proficiency-based training [220 (range, 0–452; \( P < 0.001 \)].

Conclusions: Simulator training to automaticity takes more time but is superior to proficiency-based training, as it leads to improved skill acquisition and transfer. Secondary task metrics that reflect trainee automaticity should be implemented during simulator training to improve learning and skill transfer.

Introduction

In the last decade, the traditional surgeon-training paradigm has undergone a significant shift. The Halstedian apprenticeship model seems outdated for the needs of today's surgical trainees and the demands of the current health care environment.\(^1\,^2\) In an effort to improve training, surgical educators, and national societies have embraced simulation following the example of other industries such as aviation.\(^3\) New simulators and curricula have been developed that enable training outside the operating room and enhance resident performance before patient encounters.

These efforts have been fueled by the evidence provided in the literature regarding effective transfer of simulator-acquired skill to the clinical environment.\(^5\,^6\,^9\) Furthermore, to maximize simulation training effectiveness, a proficiency-based training paradigm in which learners are required to achieve expert-derived performance goals has been suggested.\(^10\) This type of training which, according to several experts, is ideal for training on simulators, is tailored to individual needs and ensures acquisition of uniform skill by learners. Nevertheless, although proficiency-based curricula have proven to be effective in improving operative performance,\(^8\,^10\,^12\) we have clearly demonstrated before that simulator-trained learners uniformly outperform control subjects but do not reach expert performance in the operating room(OR).\(^13\,^14\,^15\) We postulate that the root of this incomplete skill transfer is that we do not reliably detect when learning is complete on the simulator because of incomplete metrics of performance, and this is unmasked in the demanding environment of the OR. Hence, although proficiency-based simulator training is effective, it may not foster optimal skill acquisition.

Most current simulation curricula use the traditional metrics of time and errors for performance feedback and assessment. Global rating scales are also being used for performance assessment but rely on the subjective opinion of the assessor and are difficult to use during simulator training, as it is not feasible to have an expert rater present during training of multiple learners. Other performance metrics such as motion recordings have also been demonstrated to be valid in distinguishing individuals with different
skill levels, but their value during simulator training is poorly understood.\(^2\) In a previous study from our group that examined the value and relationship of time and motion efficiency performance metrics during proficiency-based simulator training, we demonstrated that time was the more robust metric, as motion metrics (path length and smoothness) were easier to achieve than time by the majority of trainees.\(^6\) Importantly, the aforementioned metrics do not provide a complete picture of the attentional demands required by the primary task, the effort the performer had to invest, and the quality of the learning that occurred.\(^{17–20}\) It is well known that although 2 performers may produce equal results on time and accuracy measurements, they may have substantial differences in workload, attention demands, and physiologic parameters that reflect differences in learning, true skill level, and experience.\(^{17–20}\)

One of the main characteristics that distinguish skilled performers (experts) from novices is their ability to engage in certain activities without requiring significant attentional resources. To describe this characteristic, psychologists first used the term automaticity.\(^{21,22}\) Many habitual or highly practiced motor acts can be performed automatically, leaving enough spare attentional capacity for engagement in multiple activities. Evidence of automaticity has been used in the motor skill literature to identify skilled performers and confirm learning by novices.\(^{17–20,23,24}\) The attainment of automaticity has been mapped to specific areas of the brain that differ from those used by novices unfamiliar with a particular task.\(^{25}\) In general, automaticity is achieved through repeated practice on tasks with consistently mapped characteristics.\(^{21}\) A common procedure to measure automaticity has been the use of a secondary task that assesses spare attentional capacity when the main task is being performed.\(^{18,19,22–24}\)

We have previously used a visual-spatial secondary task for performance assessment on simulators and have demonstrated that the metrics obtained from this task were more sensitive to subtle performance differences between skilled individuals compared with the traditional metrics of time and accuracy.\(^{18}\) More recently, we also demonstrated that although novices achieve proficiency in laparoscopic suturing after relatively short training periods (based on time and errors), the attainment of automaticity (based on secondary task measures) requires significantly longer training intervals.\(^{23}\) These findings support the argument that the currently used performance metrics on simulators do not adequately reflect skilled performance and call for the incorporation of more sensitive methods such as the secondary task. In addition, they call for a study that evaluates their effectiveness in improving the incomplete transfer of simulator-acquired skill.

We, therefore, hypothesized in this study that novices who learn laparoscopic suturing on simulators would perform better in the OR after training to expert levels of secondary task performance, time, and errors (automaticity) compared with training to expert levels of time and errors alone (proficiency).

**Methods**

Our IRB-approved, single blinded, randomized controlled study was conducted at the Carolinas Simulation Center, Carolinas Medical Center, Charlotte, NC, an American College of Surgeons level-I accredited Education Institute. Eligible participants were surgical novices (premedical college students) without prior laparoscopic or simulator experience who voluntarily enrolled in the study.

**Research Design**

All enrolled participants completed a demographic and prior experience survey. After an introductory video tutorial on laparoscopic suturing and knot tying, the baseline suturing and secondary task performance of all participants was assessed on the Fundamentals of Laparoscopic Surgery (FLS) suturing model (Fig. 1). Participants were then stratified according to their performance and randomized into 2 groups in a 1:2 fashion. The first group (Control Group, n=10) did not train but participated in all testing sessions. The second group (Intervention Group, n = 20) trained in laparoscopic suturing on the FLS simulator until expert levels of time and errors (<70 second with no errors) had been achieved and were then tested together with the control group participants (Transfer test 1) on a live porcine, Nissen model. After this test, the intervention group continued training on the simulator until expert-derived performance levels on the secondary task (>73% correct detections) had been achieved in addition to the expert levels of time and errors. All participants were then retested on the porcine model (Transfer test 2).
After each training and testing session, participants completed the NASA-TLX workload assessment questionnaire. In addition, during both training and testing sessions, the heart rate and heart rate variability of participants were recorded continuously. Finally, participant performance during testing sessions was also assessed by a rater blinded to the training status of the participants using the Global Objective Assessment of Laparoscopic Skill (GOALS) assessment tool. The same rater completed all GOALS evaluations and blinding was maintained by instructing participants not to make any statements that would allow observers to infer their respective study group. Other parameters recorded during training included the number of training...
sessions, hours of training, number of repetitions, and performance scores during each repetition.

Training Details

Emphasis was given to deliberate practice; after the recommendations by Ericsson,[28] participants were trained on a well-defined task, they received feedback on their performance, and they had ample opportunities to improve their performance gradually by performing the same task repeatedly. Training session duration was kept under 1 hour,[28] and multiple training sessions during the same day were not allowed. Scheduling training was flexible and tailored to the subject's needs. Participants scheduled their own dates and times of training based on simulator and instructor availability. Instructors supervised training and provided augmented feedback to facilitate skill acquisition.[29,30] In addition, free access to video tutorials describing the suturing technique and common pitfalls and helpful tips and tricks were provided. To ensure uniform testing conditions and minimize the cost of the transfer tests, all trained participants who had achieved simulator proficiency more than 2 weeks before the transfer test had "refresher" training during which they had to demonstrate their proficiency on the simulator again.

Performance Metrics

An objective score was calculated for suturing performance assessment, both on the simulator and on the porcine model, based on task duration and errors according to a previously published formula: 

\[ \text{SCORE} = 600 - \text{COMPLETION TIME} - [10 \times \text{KNOT ACCURACY AND SECURITY ERRORS}] \]  

[14] Expert score was set at 530 (ie, task time of 70 seconds with no errors) based on prior work from our group.[18]

The secondary task used in this study was a visual-spatial processing task that addressed both perceptional and memory resources. Participants had to detect randomly appearing 4 × 4 cm squares on a laptop screen placed adjacent to the laparoscopic monitor and respond when a succession of 3 squares was identified on the right side of the screen by pressing a foot pedal. The validity of this task in distinguishing individuals of different skill levels has been demonstrated previously by our group.[18,23] In addition, another more recent study that used a similar secondary task also found it to be valuable.[24] The expert level for the secondary task was set at 73% correct detections based on the findings of our previous study.[18]

Heart rate and heart rate variability were used to monitor autonomic nervous system responses as an indicator of stress of the subjects during training and testing sessions. This physiologic measure has been shown previously to accurately reflect learner stress.[31–33] The tetherless ergonomics workstation developed by Smith and colleagues was used for the recordings.[34]

This instrument is designed to measure the physical workload and stress level of surgeons performing minimally invasive surgery, although leaving the surgeon unencumbered, thereby allowing completion of studies in complex and realistic settings.[34]

The NASA-TLX questionnaire is a workload assessment instrument, which has been used extensively and shown to be sensitive and diagnostic in aviation tasks.[19,26] We have also recently demonstrated its validity for workload assessment during training on laparoscopic simulators.[35] Subjects rate the mental, physical, and temporal demands of the task, and their performance, effort, and frustration level. Typically, the workload associated with executing a task decreases as experience with the task increases.[19,35] This metric was used to assess the workload that participants perceived to achieve a particular level of performance.

The Global Objective Assessment of Laparoscopic Skill is a global rating scale for the evaluation of laparoscopic skills of surgery residents by experienced observers. The instrument has been shown to be valid and reliable.[27]

Inadvertent injuries were defined as injuries caused by the needle or the instruments to structures other than the premarked areas of the gastric fundus (typically the esophagus, liver, remote areas of the stomach, or the diaphragm). Injuries consisted of needle sticks of these structures or liver lacerations caused by the instruments.

Simulator and Suture Model

The validated FLS simulator[5,36] that is coendorsed by SAGES and ACS was used for simulator training. To accomplish the suturing task, trainees had to place a 2 to 0 silk suture through the premarked targets of the FLS model and tie intracorporeal knots (1 surgeon's and 2 square knots) using conventional laparoscopic instruments (ENDOPATH Needle Holder, Ethicon Endo-Surgery Inc). The maximal allowable task time was 10 minutes, and the knots were assessed for accuracy and security, as has been described previously.[14]

Porcine Testing Model
The laparoscopic Nissen fundoplication model has been used in several prior studies for performance assessment in the OR after prior training on simulators. This model offers the advantage of assessing performance under real OR conditions. It allows standardization of testing sessions, makes the use of a single animal for numerous individuals possible, realistically replicates the human's working space constraint and diaphragmatic motion artifact, and alleviates the cost, ethical, and logistical concerns with performing a transferability trial on humans. Participants were required to perform the same task they learned on the simulator; however, instead of the FLS model, they placed sutures at premarked areas of the gastric fundus to create a fundoplication. Maximal allowable task time was 10 minutes, and their knots were objectively assessed for accuracy and security.

**Sample Size Calculation and Data Analysis**

The sample size of this study (n = 30) was chosen to allow the detection of at least a 20% performance difference between the groups during testing sessions with a power of 0.8 and an alpha level of 0.05. It was based on the performance achieved by a group of novices in the porcine model after simulator training in laparoscopic suturing on a prior study from our group. The Sigma Stat software was used for statistical analysis (SigmaStat version 3.0, SPSS Inc, Chicago, IL). The Kolmogorov-Smirnov test was used to assess data normality. Data are reported as mean ±SD or median (range) based on their normality. Repeated measures ANOVA or repeated measures ANOVA on Ranks as indicated was used to compare intragroup performance differences between baseline, OR test 1 and OR test 2. The unpaired t test or Signed Rank test as indicated was used for intergroup comparisons. A P-value of <0.05 was considered significant.

**Results**

Thirty participants voluntarily enrolled in the study protocol. There were no significant differences in the demographics, prior surgical and simulator experience, and simulator performance between groups at baseline (Table 1).

### Table 1. Participant Baseline Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Control Group N = 10</th>
<th>Intervention Group N = 20</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>25 ±2</td>
<td>24 ±3</td>
<td>0.84</td>
</tr>
<tr>
<td>Sex (% women)</td>
<td>50</td>
<td>45</td>
<td>0.52</td>
</tr>
<tr>
<td>Handedness (% right)</td>
<td>90</td>
<td>85</td>
<td>0.65</td>
</tr>
<tr>
<td>Prior simulator experience (hours)</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Prior surgical experience (No. of observed cases)</td>
<td>0 ±3</td>
<td>0±4</td>
<td>0.72</td>
</tr>
<tr>
<td>Prior video game experience (1–10 Likert scale)</td>
<td>6 ±3</td>
<td>7±3</td>
<td>0.67</td>
</tr>
<tr>
<td>Baseline heart rate</td>
<td>104 ± 14</td>
<td>99 ± 15</td>
<td>0.35</td>
</tr>
<tr>
<td>Baseline suturing scores</td>
<td>16 ± 32</td>
<td>6 ± 25</td>
<td>0.23</td>
</tr>
<tr>
<td>Baseline secondary task scores (% correct detections)*</td>
<td>98 ±2</td>
<td>97 ±3</td>
<td>0.89</td>
</tr>
</tbody>
</table>

*When performed as primary task.

Three intervention group participants (15%) discontinued their study participation after their initial practice session. All others (n = 17) achieved simulator proficiency on the suturing task after 54 ± 14 repetitions and participated in the first OR test. After the first OR test, 7 participants (41%) were able to achieve secondary task goals concurrently with suturing proficiency (automaticity) after an additional 109 ± 57 repetitions. Two participants discontinued their participation during this second phase of training. One year after the study began, all active participants were asked to take the second OR test. Twelve participants from the intervention group and 8 controls completed the second OR test. Three intervention group and 2 control group participants could not participate because of scheduling conflicts (ie, total study attrition rate was 33%).

Intervention group participant suturing performance improved significantly after proficiency-based training compared with their baseline and remained near expert levels during automaticity training without significant changes (Fig. 2). On the contrary,
secondary task scores of the intervention group participants changed minimally between the beginning of training and achievement of proficiency (\(P = \text{n.s.}\)) but improved significantly by the end of the study protocol (correct detections baseline 0 ± 0 vs before OR test 1, 1 ± 1 vs before OR test 2, 12 ± 7; \(P < 0.001\)) (Fig. 2).

![Image](https://www.medscape.com/viewarticle/757914_print)

**Figure 2.** Intervention group primary and secondary task performance during simulator training. Although suturing performance improved significantly during initial training to proficiency and then plateaued, secondary task performance improved mainly during automaticity training.

Control group performance did not change significantly between baseline and the OR tests for any of the recorded parameters. On the contrary, the intervention group demonstrated significantly better performance during both OR tests compared with their baseline (Fig. 3). Importantly, during the second OR test, they achieved better suturing performance scores [345 (range, 0–537) vs. 220 (range, 0–452); \(P < 0.001\), Fig. 3] and GOALS ratings 22 (range, 16–25) versus 16 (range, 13–21); \(P = 0.003\), Fig. 4) compared with their first OR test. Notably, the intervention group caused fewer inadvertent injuries during their second OR test compared with their first OR test [1 (range, 0–3) versus 4 (range, 0–13); \(P < 0.001\), Fig. 5].
Figure 3. Objective suturing scores achieved by groups at baseline and during OR tests. *P*-values reflect differences between groups at each interval. Box plots and error bars represent medians and 10th, 25th, 75th and 90th percentiles. Dots represent 5th and 95th percentiles. Performance during both OR tests was significantly better compared with their baseline for the intervention group (*P* < 0.001) but not for the control group. In addition, intervention group performance during OR test 2 was significantly better than OR test 1 (*P* < 0.001).
Figure 4. Group comparison of GOALS scores during OR tests. Box plots and error bars represent medians, 10th, 25th, 75th and 90th percentiles. GOALS scores differed significantly between groups during both OR tests ($P < 0.001$).
Subgroup analysis that compared the 7 participants who reached secondary task expert levels with the 5 participants who did not reveal no differences at baseline but better performance of those who reached automatcity during both OR tests (Table 2). Nevertheless, these differences approached but did not reach significance ($P = 0.07–0.11$) likely due to the small sample size. On the contrary, both groups performed better during OR test 2 compared with their OR test 1 performance (Table 2; $P < 0.05$ for both groups). Importantly, the 7 participants who achieved automaticity clearly outperformed the controls during OR test 2 [363 (range, 158–537) versus 0 (range, 0–385), respectively, $P < 0.001$]. However, even those who did not reach the expert level on the secondary task, still performed significantly better than the controls ($P < 0.001$).

**Table 2. Subgroup Analysis of Intervention Group Participants Based on Their Ability to Reach Expert Levels of Secondary Task Performance During the Study Period**

<table>
<thead>
<tr>
<th></th>
<th>Reached automatically</th>
<th>Did not reach automaticity</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suturing scores at baseline</td>
<td>0 ±0</td>
<td>0±0</td>
<td>1</td>
</tr>
<tr>
<td>No. of repetitions to reach proficiency</td>
<td>50 ± 11*</td>
<td>63 ± 15</td>
<td>0.12</td>
</tr>
<tr>
<td>Suturing scores at proficiency</td>
<td>552 ± 12</td>
<td>545 ±9</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Heart rate recordings revealed that although the heart rate of the control group participants did not change significantly between baseline on the simulator (100.5; range, 94.5–109) and OR test 1 (100; range, 99–107) versus OR test 2 (102.5; range, 97.5–116), $P > 0.05$, it was significantly higher during the OR tests for the intervention group participants compared with their baseline (Fig. 6). In addition, although the median heart rate was not statistically significantly different between groups at any time point, the change in heart rate between baseline and OR tests 1 and 2 was. In particular, the change in HR for the control group versus the intervention group between baseline and OR test 1 was $-2.5$ beats/min (range, $-30$ to 33) versus 11 ($-12$ to 38), respectively ($P < 0.05$) and between baseline and OR test 2 1.5 ($-17$ to 16) versus 16 ($-6$ to 29), respectively ($P < 0.01$). No differences were noted in the HR changes of the groups between OR test 1 and 2. No differences in heart rate variability were found between groups or between the baseline and OR tests for the intervention group.

<table>
<thead>
<tr>
<th></th>
<th>Reached automaticity</th>
<th>Did not reach automaticity</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suturing scores at baseline</td>
<td>0 ±0</td>
<td>0±0</td>
<td>1</td>
</tr>
<tr>
<td>No. of repetitions to reach proficiency</td>
<td>50 ± 11*</td>
<td>63 ± 15</td>
<td>0.12</td>
</tr>
<tr>
<td>Suturing scores at proficiency</td>
<td>552 ± 12</td>
<td>545 ±9</td>
<td>0.25</td>
</tr>
<tr>
<td>Secondary task scores prior to OR test 1 (%)</td>
<td>5 ±6</td>
<td>0±1</td>
<td>0.21</td>
</tr>
<tr>
<td>Suturing scores during OR test 1</td>
<td>249 ± 165</td>
<td>154 ± 158</td>
<td>0.07</td>
</tr>
<tr>
<td>No. of total repetitions</td>
<td>175 ± 71*</td>
<td>139 ± 52</td>
<td>0.35</td>
</tr>
<tr>
<td>Secondary task scores prior to OR test 2 (%)</td>
<td>53 ± 15</td>
<td>20 ± 19</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Suturing scores during OR test 2</td>
<td>351 ± 117</td>
<td>283 ± 128</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Participant performance during OR test 2 was better as compared to OR test 1 for both groups ($P < 0.05$).

* $P < 0.01$.  

Table 2. Subgroup Analysis of Intervention Group Participants Based on Their Ability to Reach Expert Levels of Secondary Task Performance During the Study Period
Figure 6. Heart rate changes during transition from simulator to the OR. The change in heart rate between baseline and OR tests was significantly different between the intervention and the control groups ($P < 0.05$).

Participant workload also did not change significantly between baseline and OR tests 1 and 2 in the control group. In contrast, workload scores decreased after training to proficiency on the simulator, increased back to baseline levels during OR test 1, dropped to end of proficiency levels again during simulator training to automaticity but did not rise again during OR test 2 (Fig. 7). In addition, the intervention group NASA scores during OR test 2 were significantly lower compared with OR test 1 (60 ± 30 vs 89 ± 16.5; $P = 0.01$).
Figure 7. Workload comparison between groups at different time intervals. P-values refer to changes in NASA scores between depicted time intervals. No differences in workload were noted for the control group at all intervals. The intervention group workload scores were lower during OR test 2 compared with OR test 1 and compared with control group workload scores during OR test 2. There are fewer data points for the control group, as these participants did not train and only participated during the testing sessions.

Discussion

The results of this study confirmed our hypothesis that the transfer of simulator-acquired laparoscopic skill to the clinical environment can be augmented by the incorporation of a secondary task into the curriculum. We showed that training novices to expert levels of secondary task performance above and beyond the standard proficiency levels of time and errors (automaticity) leads to superior skill acquisition and transfer compared with traditional proficiency-based training alone. In addition, we demonstrated that the achievement of automaticity on simulators requires an extensive amount of training well beyond what is needed to reach proficiency as we define it today. In fact, on the basis of the number of repetitions, study participants who achieved automaticity did so after training that was twice as long as their initial training to proficiency. Moreover, at least half of the participants were not able to achieve expert levels of secondary task performance despite prolonged training. On the contrary, they all demonstrated improvements in their secondary task performance indicating that even longer training intervals were necessary for the participants who failed to achieve automaticity during the study period. This finding confirms the results of a previous study from our group demonstrating that the achievement of automaticity requires significantly longer training intervals than training to proficiency.[23]

It is also important to note that our participants not only were able to achieve better performance after automaticity training but also...
were safer in the OR as they caused significantly fewer inadvertent injuries to surrounding structures compared to initial proficiency training. Moreover, the finding that the injuries caused by the trained group during the first OR test were significantly more than those of the control group indicates that they were not ready for the transition to the clinical environment after initial proficiency training. The higher level of inadvertent injuries seen in the intervention group after initial proficiency training may be the result of a speed-accuracy tradeoff.\(^\text{[17,37]}\) It is possible that participants likely elected to perform the task as quickly as possible compromising accuracy and safety for speed. This finding is disconcerting and warrants further investigation, as most currently available skills curricula are proficiency-based. Emphasis on accuracy rather than speed during training may be needed. Another explanation may be tied to the increased stress that we observed in our participants based on their heart rate recordings during both OR tests. Unlike the control group participants, who did not demonstrate appreciable changes in their heart rates between baseline and OR tests, intervention group participants had significantly higher heart rates during the OR tests indicative of their stress. We have previously published similar findings regarding the transition of novice learners from the simulator to the OR and have speculated that it is a consequence of performance anxiety.\(^\text{[15]}\) Accordingly, it seems that intervention group participants may have experienced higher levels of stress when they realized that the OR task is significantly more difficult than the simulator task and that their OR performance is poorer than they anticipated after having reached expert levels on the simulator. This does not seem to apply to the controls as the lack of prior simulator training eliminated a comparison experience for their poor OR performance. Furthermore, recall that the suture score for the control group was nearly zero. Thus, it is likely that the controls did not cause many injuries because they were not competent enough to even perform the suturing. Nevertheless, it seems that after automaticity training, participants in the intervention group were able to better handle their stress, as it did not influence their performance to the degree it did during the first OR test. Given the known detrimental effects of stress on performance,\(^\text{[15]}\) our findings support the need for stress coping strategies that will help minimize performance deterioration during stressful situations often encountered in surgery.

Additional evidence for the improved performance of our participants after automaticity training comes from the NASA-TLX workload ratings. These ratings increased dramatically during the first transition to the OR from the simulator but did not change significantly during the second transition. This indicates that the participants found that performing the suturing task on a live animal was significantly more difficult compared with the simulator after proficiency training, but not after automaticity training. In other words, extended training on the simulator made performance of the task in the clinical environment easier. The validity of the workload ratings is supported by a recent publication from our group that demonstrated across several studies that trainee workload changes on simulator's mirror performance changes; the poorer the performance the poorer the workload ratings and vice versa.\(^\text{[33]}\)

These findings suggest that, to maximize skill transfer to the OR, simulator training should go well beyond initial achievement of proficiency, and the use of a secondary task that documents achievement of automaticity should be incorporated into the skills curriculum.

Overlearning refers to deliberate practice on a task beyond a set criterion performance level and is known to be a very important determinant of skill acquisition and retention.\(^\text{[12,17]}\) Nevertheless, the optimal amount of overtraining is task specific, and excessive training can lead to skill degradation especially for simpler tasks.\(^\text{[17]}\) For a complex task such as intracorporeal suturing and knot tying, training past initial acquisition of proficiency is likely necessary to ensure long-term retention. The secondary task metric may prove very useful in identifying the ideal amount of overlearning for a particular task, as it documents when learning is sufficient. Along these lines, our findings indicate that the amount of overlearning necessary for the laparoscopic suturing and knot-tying task is double that of the training amount needed to achieve initial proficiency.

Decrement observed in the secondary task under dual-task conditions (primary + secondary tasks) compared with the secondary task performed by itself reflect the mental demand of the primary task. These conditions are met when both tasks compete for the same pool of resources. We elected to use a visual-spatial secondary task instead of a simpler task (such as auditory or mathematical) for several reasons. Secondary task measures should satisfy several important criteria.\(^\text{[38,39]}\) First, they must show sensitivity (ie, they must reflect changes in task difficulty or resource demand). Second, they must be selective (ie, they should reflect differences in workload and resource demand related to the primary task, but unrelated factors such as physical workload). Third, they should be unobtrusive (ie, they should not interfere with performance on the primary task whose workload is under assessment). In a comprehensive review of the properties of many secondary tasks, Lysaght and his colleagues\(^\text{[40]}\) found that few secondary tasks satisfied all criteria. Thus, the choice of a secondary task should be dictated by the nature of the primary task and secondary task characteristics that satisfy as many criteria as possible. Specifically, a visual-spatial primary task (eg, driving) paired with a visual-spatial secondary task (eg, looking for a specific street sign) will reflect differences in mental workload between different driving conditions (eg, slow moving light traffic and fast moving heavy traffic) or between novice and experienced drivers.
By contrast, Eggemeier[39] argues that mismatches in the resources demanded by the primary and secondary tasks will not reveal differences between single and dual-task performance. Thus, a visual-spatial primary task (eg, driving) paired with an auditory or verbal secondary task (eg, listening to the radio or carrying on a conversation) may not reflect any differences in mental workload. In fact, most people can drive and listen to the radio or carry on a conversation with little difficulty even under very challenging driving conditions. Support for this approach can be found in O’Donnell and Eggemeier[38] and Vidulich.[41] Thus, given that laparoscopy places significant demands on visual-spatial attention, we elected to use a visual-spatial secondary task to compete for the same resources to maximize the sensitivity of our secondary task to detect changes in mental workload.

Limitations of our study include the high attrition rate of our participants (33%) and the additional compromise of our sample size by the unavailability of some of our participants to take the second OR test. Nevertheless, our attrition rate is not unexpected for a training study that extends over such a long time period. Furthermore, a post hoc power analysis based on the actual suturing performance (primary outcome) differences between OR test 1 and 2 and our effective sample size (n = 12) revealed that our power was preserved at 0.81 (with a Cohen’s d effect size of 0.77). Our study power was preserved because the actual performance difference between the 2 tests was 54%, far higher than the threshold of 20% we had chosen during our original power calculation. Another limitation of this study is that many intervention group participants (41%) did not achieve expert levels of secondary task performance before OR test 2. Nevertheless, we elected to perform the second OR test before everyone achieving automaticity, because the training duration became too lengthy, the study budget was running low, and we were concerned about even higher attrition rates. In addition, even those participants who did not achieve expert levels of secondary task performance within the study duration demonstrated improvements in secondary task scores and significantly improved performance during OR test 2 compared with OR test 1 clearly benefitting from the additional training. Importantly, if all participants were to have reached expert levels of secondary task performance, we would likely have observed a more profound effect of training on performance. It should also be noted that this was a single institution study, which may have introduced bias in its results; however, we believe we have minimized this effect by using a randomized and blinded study design.

In conclusion, simulator training to automaticity takes more time but is superior to proficiency-based training, as it leads to improved skill acquisition on simulators and transfer to the clinical environment. Metrics obtained from a secondary task provide a more comprehensive assessment of trainee performance and multitasking ability compared with the traditional metrics of time and errors and should be implemented during simulator training to improve learning. Additional, multi-institutional studies are needed to confirm these findings.

References

41. Vidulich MA. Mental workload and situation awareness: essential concepts for aviation psychology practice. In: Tsang PS,
Acknowledgments
The authors thank Drs. Daniel J. Scott, James R. Korndorffer Jr., B. Todd Heniford, Ramon Berguer, H. James Norton, Ajita S. Prabhu, Yuliya Y. Yurko, and Phuong H. Nguyen who contributed during various phases of this study. They also thank Cissy Swartz for the editing of this manuscript.