ORIGINAL ARTICLE

ONLINE FIRST

Criterion-Based (Proficiency) Training to Improve Surgical Performance

Marvin P. Fried, MD; Rachel J. Kaye, MD; Marc J. Gibber, MD; Alexis H. Jackman, MD; Boris P. Paskhover, MD; Babak Sadoughi, MD; Bradley Schiff, MD; Rebecca E. Fraioli, MD; Joseph B. Jacobs, MD

Objective: To investigate whether training otorhinolaryngology residents to criterion performance levels (proficiency) on the Endoscopic Sinus Surgery Simulator produces individuals whose performance in the operating room is at least equal to those who are trained by performing a fixed number of surgical procedures.

Design: Prospective cohort.

Setting: Two academic medical centers in New York City.

Participants: Otorhinolaryngology junior residents composed of 8 experimental subjects and 6 control subjects and 6 attending surgeons.

Intervention: Experimental subjects achieved benchmark proficiency criteria on the Endoscopic Sinus Surgery Simulator; control subjects repeated the surgical procedure twice.

Main Outcome Measures: Residents completed validated objective tests to assess baseline abilities. All subjects were videotaped performing an initial standardized surgical procedure. Residents were videotaped performing a final surgery. Videotapes were assessed for metrics by an expert panel.

Results: Attendings outperformed the residents in most parameters on the initial procedure. Experimental and attending groups outperformed controls in some parameters on the final procedure. There was no difference between resident groups in initial performance, but the experimental subjects outperformed the control subjects in navigation in the final procedure. Most important, there was no difference in final performance between subgroups of the experimental group on the basis of the number of trials needed to attain proficiency.

Conclusions: Simulator training can improve resident technical skills so that each individual attains a proficiency level, despite the existence of an intrinsic range of abilities. This proficiency level translates to at least equal, if not superior, operative performance compared with that of current conventional training with finite repetition of live surgical procedures.


Author Affiliations:
Department of Otorhinolaryngology–Head and Neck Surgery, Montefiore Medical Center, Bronx, New York (Drs Fried, Kaye, Gibber, Jackman, Sadoughi, Schiff, and Fraioli); Department of Otolaryngology, Yale University School of Medicine, New Haven, Connecticut (Dr Paskhover); and Department of Otolaryngology, New York University Medical Center, New York (Dr Jacobs).

Technical abilities are highly individual, as shown by the wide range of ability characterizing different musicians, artists, and surgeons. Given that the issue of creating a competent and safe surgeon is of paramount importance, we hypothesize that the objective measurement of a resident’s progress is critical to the achievement and the assessment of proficiency. We are using the term proficient here with the intention of differentiating it from competent. The Oxford English Dictionary Online defines proficiency as “a skill, a talent; a certain standard of skill acquired after a period of education or training” whereas competence is defined as “sufficiency of qualification; capacity to deal adequately with a subject.” Thus, proficiency applies to the attainment of a particular technical skill and competence designates the ability to manage all aspects of a particular subject. The acquisition of such skills is of vital importance during residency training. To this end, an innovative state-of-the-art simulation device, the Endoscopic Sinus Surgery Simulator (ES3), was developed by defense contractor Lockheed Martin, Inc.

The ES3 has passed through a rigorous validation process to establish face and content validity,1 construct validity to discriminate between user experience,2,3 the potential to become a surgical trainer by increasing surgical confidence,4 and predictive validity for the transfer of simulator-acquired skills to operative procedures.5 It has been previously shown that training on the simulator improves mean per-
formance (to a level of proficiency) and consistency. The ES3 ensures skill proficiency by training to expert-level criteria. This is in contrast to current surgical training that typically lasts for a specified period or number of procedures, thereby producing surgeons with variable skill levels, a practice that is becoming less acceptable for patient safety. Indeed, residents are required to complete a specific number of procedures in various subspecialty areas, with proficiency assumed and not measured. We propose to use the ES3 to train otorhinolaryngology residents to criterion performance levels and to investigate whether this criterion-based training produces residents whose performance in the operating room is at least equal if not superior to those residents who are trained by performing a fixed number of procedures. Although the field of general surgery has compared conventional training with proficiency-based simulator training, the conventional training is often poorly defined and no finite number of task repetitions (operative procedures) is described. To our knowledge, this study is the first such assessment within the field of otorhinolaryngology and it does so through a rigorous method.

This study used 2 accredited, university-based programs in New York City as its site locations. After obtaining institutional review board clearance from both participating institutions, 14 junior (postgraduate year [PGY] 1-3) otorhinolaryngology residents and 6 attending surgeons served as subjects (ATT group). The experimental (EXP) and control (CTRL) groups were defined as junior otorhinolaryngology residents who would be either trained to proficiency on the simulator (EXP group) or trained in the standard fashion by performing a limited number of defined sinus surgery procedures (CTRL group). Subjects were selected for each group primarily on the basis of site and, as such, they were evaluated for equivocal fundamental abilities. Of note, the relatively small number of residents per PGY in otolaryngology training programs in general, as well as specifically at our 2 site locations, was a considerable obstacle in obtaining subjects to constitute our EXP and CTRL groups despite aggressive recruiting. Moreover, eligible surgical cases were limited because most endoscopic sinus surgery (ESS) cases were unable to fulfill our strict inclusion criteria.

**SUBJECT INCLUSION/EXCLUSION CRITERIA**

All residents were in PGY 1 to 3. Exclusion criteria were based on prior experience to assure inexperience with ESS, defined as having performed fewer than 5 cases as the primary surgeon. All potential residents were filtered by a preliminary questionnaire (Figure 1) to establish whether they fulfilled the inclusion criteria and all potential subjects were found to be eligible. The preliminary questionnaire also assessed whether there was any gross cause of possible fine motor skill disparity between the EXP and CTRL groups.

**RECRUITMENT OF SURGICAL PROCEDURES**

Stringent criteria were established regarding case eligibility and included the following:

1. **Inclusion criteria**
   - Must have an appropriate indication for ESS
   - Must be older than 18 years (either sex)
   - No racial, ethnic, or religious restrictions
   - American Society of Anesthesiologists physical status classification system designation of 1 to 3

2. **Exclusion criteria**
   - Positive bleeding history
   - Prior ESS
   - Gross pathology that obstructs visualization of the target areas of operation (eg, massive polyposis, severe septal deviation)

**DATA COLLECTION/STUDY DESIGN**

After designation of resident subjects into either an EXP or a CTRL group as previously defined, all resident subjects completed the preliminary questionnaire and 3 test trials on the novice mode of the ES3 (Figure 2). Proficiency criteria for the abstract environment of the novice mode have been previously established using attending surgeons as benchmarks. All resident subjects were then videotaped while performing a surgical procedure (including navigation, injection, uncinectomy, and maxillary antrostomy) on patients. The surgical tasks that were completed with each videotaping session were previously defined and standardized (Figure 3). Both resident groups were then “trained” in different manners. The CTRL group assisted in 2 additional ESS cases and performed the same surgical procedures (navigation, injection, uncinectomy, and maxillary antrostomy) without being videotaped. The EXP group performed trials in the intermediate mode of the ES3 until reaching benchmark proficiency criteria previously described (3 consecutive trials of >93.9). All resident subjects were then videotaped while performing the same surgical procedure (navigation, injection, uncinectomy, and maxillary antrostomy) on patients. The ATT group established benchmark criteria and were videotaped while performing the same surgical procedures as the resident subjects. All subjects were deidentified in the videotapes and their anonymity was preserved throughout the study.

**ESTABLISHMENT OF A RATER PANEL**

Three senior academic otorhinolaryngologists with expertise in ESS were enlisted independently and asked to establish interrater reliability. Consistency among the raters was achieved.
by reviewing selections of sample videotaped ESS procedures that were unrelated to the studied population. The expert panel was able to standardize their scoring system with these sample videotapes. This expert panel was masked to subject and rated the videotaped recordings using a previously described custom-made software application that incorporates consensual ESS metrics. The measured variables were time to completion of task, case difficulty, tool manipulation, tissue respect, task completion rate, surgical confidence (10-point scale), and number of errors. All these parameters were applied to the review of the 3 main recorded tasks.

RATING THE VIDEOTAPED INTRAOPERATIVE PERFORMANCES

The masked expert panel rated the videotaped recordings using a previously developed software application named the Universal Rating and Assessment Tool for Examiners. This is an interactive examination tool that is unique and was designed to assess the videotaped procedures. The rater is able to select relevant screen buttons displaying error types in real time as the errors are observed. A questionnaire window prompts the rater to grade the difficulty, tool manipulation, tissue respect, completion, and confidence of each videotaped segment via a numerical scale with a range of 1 to 10.

RESULTS

Statistical significance was defined as $\alpha < .05$. Results of the preliminary questionnaire (Figure 1) were analyzed using the Fisher exact test. There was no statistical significance between the EXP and CTRL groups for any of the questions except for PGY ($P = .02$). The CTRL group was composed entirely of residents in PGY 2, while the EXP group was composed of residents in PGY 1 (25%), PGY 2 (25%), and PGY 3 (50%); the mean (SD) EXP group PGY was 2.25 (0.83). However, we do not believe that this is clinically significant because the resident subject pool did not differ in any other variable and, most important, all resident subjects had performed fewer than 5 ESS cases as the primary surgeon. In addition, there was no difference between EXP and CTRL groups on the initial assessment tests on the ES3 or in the initial videotaped procedure in the operating room. Thus, we found no gross cause for fine motor skill discrepancy between the groups. Results of the initial ES3 assessment were analyzed by performing a t test for each of the 3 trials, comparing CTRL and EXP groups. No significant difference was found between the groups, showing that their fundamental ability was similar through an objective skill parameter. The number of trials necessary for EXP subjects to achieve proficiency on the intermediate mode of the ES3 was recorded and analyzed for all 8 subjects. The mean (SD) number of trials was 9.12 (3.7), with 5 subjects needing fewer than 6 trials and the remaining 3 subjects requiring more than 12 trials.

In regard to the rater panel, interrater reliability was assessed using the Cohen weighted $k$ test by creating a single variable for each rater ($n=310$) for all 102 video segments with 5 parameters (difficulty, tool manipulation, tissue respect, task completion, and surgical con-
The Cohen weighted \( \kappa \) test showed that raters 1 and 2 had fair agreement (\( \kappa = 0.2299 \)), whereas raters 1 and 3 as well as 2 and 3 had slight agreement (\( \kappa = 0.182 \) and 0.138, respectively). However, we noted that the average numerical score for all observations for raters 1, 2, and 3 were 5.19, 5.38, and 6.94, respectively (range, 1-10). This data suggested that rater 3 had a possible “leniency bias.” In addition, the accuracy of raters was based on the ability to differentiate between the initial videotaped procedure of the ATT group and that of the novice surgeons. However, rater 3 did not sufficiently differentiate these 2 groups on analysis of variance. Thus, rater 3 had suboptimal accuracy and was excluded from further analysis. The data stemming solely from raters 1 and 2 showed extreme uniformity of significance on most parameters that we assessed and thus was sufficient for analysis.

The difference in numerical scores between the initial and final videotaped procedures for the EXP and CTRL groups was analyzed by paired \( t \) tests for all resident subjects. There was no significant difference in the difficulty level of the initial and final procedures for the EXP and CTRL groups. Both EXP and CTRL group subjects had statistically significant superior performance on their final procedures (compared with initial procedures) for all parameters except for tool manipulation in navigation tasks (Figures 4, 5, and 6). The \( t \) tests comparing the CTRL and EXP groups in the initial and final videotapes for all tasks and procedure types showed that there was no significant difference in their initial videotaped procedure scores. However, the EXP group had statistically significant superior scores in all of the final navigation tasks.

We then used analysis of variance to compare the ATT, CTRL, and EXP groups for all parameters. Of note, neither the initial nor final procedures differed significantly in case difficulty. There was a statistically significant difference in all characteristics on the initial procedure (except for tissue respect in initial dissection), with the ATT group scoring superiorly. This demonstrated that although there was no difference in case difficulty, a proficiency level achieved by ATT subjects is identifiable by rating videotaped procedures using the defined metrics. In the final navigation tasks, there was a statistically significant difference between the groups for tissue respect and confidence. The EXP and ATT subjects had equivocal scores on the final navigation procedure, with the CTRL group having significantly lower scores for all described characteristics (Figure 4). Thus, EXP subjects per-
formed significantly superior to CTRL subjects on the final navigational tasks and, in fact, were indistinguishable from ATT subjects in their performance of these tasks. Although the EXP subjects had superior scores compared with the CTRL subjects for all of the final injection tasks, the differences were not significant. In the final dissection tasks, the ATT group had statistically significant superior scores on completion and confidence.

Returning to the different number of trials needed to achieve proficiency levels on the simulator, we noted that the EXP group could be divided into 2 subgroups: those who achieved proficiency levels on the intermediate mode of the simulator in fewer than 6 trials with 5 subjects and those that achieved such proficiency in more than 12 trials with 3 subjects. The t test analysis showed a significant difference between these subgroups on the tool manipulation parameter during the videotaped initial navigation procedure. This can best be explained by the small number of subjects within each subgroup. Of greater importance is the comparison of the subgroup performance on all other tasks, especially the final dissection, of which there were no statistically significant differences.

**COMMENT**

At the core of surgery lies the ability and skill to perform operations intrinsic to the specialty, and this is intimately related to patient safety in the operating room. Currently, there is no objective compulsory technical skill assessment of surgical residents, and therefore the training process that molds a resident into a surgeon is subjective, as is the continuing evaluation of skill, dexterity, and insight. The subjective and objective assessment of skills is extensively studied in the laparoscopic surgery simulation literature, in which the classic adage of “see one, do one, teach one” is becoming increasingly ineffective in training because of growing medicolegal concerns for patient safety and new pressure toward improving operating room efficiency. Surgical simulators seek to fill the need for adjunctive training outside of the operating room, a resource made increasingly imperative by the growing time and financial limitations of resident intraoperative training. Simulation has emerged as a technological advancement purported to provide an objective measurement of resident training in order to achieve skill proficiency established by benchmark criterion. The acquisition of surgical skills currently requires live patient experience with finite availability; this produces a wide range of abilities within the resident population. Because not all trainees develop skill sets equivalently, criterion-based training seeks to reduce this variability in skill level and ensure that trainees have at least reached a level benchmarked by experienced and safe surgeons. Furthermore, leaving resident education in the hands of finite encounters is becoming increasingly unacceptable, and surgical practice lends itself to a particular need to train residents to a level of competence. In his presidential address at the Central Surgical Association’s annual meeting, Richard Bell recommended surgical simulation as the solution to the finite and “inadequate” level of operative experience available during residency training. Internationally, the increasing use of virtual reality (VR) simulators is apparent because proficiency-based training programs with VR simulators are more readily available and the potential of simulation being used for certification has been suggested. This novel concept has already begun to impact resident credentialing in the United States; the American Board of Surgery has required the Fundamentals of Laparoscopic Surgery (which includes hands-on simulation exercises) for certification since 2009.

The present data displays significant uniformity within multiple parameters, yielding 2 major outcome measures. The first major outcome is the initial learning curve, it contains the potential to protect patients from novice errors during this period.

The second major outcome regarded the ability of VR-trained residents to be trained to criterion levels, consistently demonstrating a level of proficiency. We established that a disparity existed with our resident subjects regarding the number of trials needed to train to proficiency on the ES3. This disparity represents the wide range of aptitudes normally present in a resident population. It has been shown in multiple studies that the number needed to attain competency is variable within the resident population owing to individual learning differences within the group. Proficiency-based training seeks to eliminate this variability by ensuring that participants achieve validated criteria. We also demonstrated that once a subject attained proficiency on the ES3, whether it took that individual fewer than 6 training sessions or more than 12, their surgical performance was indistinguishable. Thus, proficiency on the ES3 directly translates to proficiency with surgical skills regardless of the number of trials needed to attain that proficiency level. This is in direct contrast to traditional training with a finite number of cases, which can lead to variably trained individuals. Thus, the potential truly exists to have variable case volumes that are student dependent, not predetermined by an external panel or board. Furthermore, proficiency-based training could be implemented as an independent study approach that is tailored to each individual.

This study is unique in that it seeks to evaluate an alternative to the cornerstone of early surgical education—technical experience with live patients. One of the benefits of the current method is that the presence of multiple videotaped procedures for each subject allows us to determine not only the subjects’ objective performance at
a fixed point in time during their training but also their improvement. However, the study was limited in the number of subjects per training group. This was due to the relatively small number of residents per PGY in otolaryngology training programs. Moreover, eligible surgical cases are restricted because most ESS cases are unable to fulfill our rigorous inclusion criteria.

In conclusion, this study demonstrates that simulator training can improve resident skills to reach a proficiency level for all members of the group, despite the existence of an inherent range of abilities. Furthermore, we have demonstrated that this proficiency level translates to equal, if not superior, operative performance when compared with that of conventional training with finite repetition of live surgical procedures. This shows that proficiency training catapults the novice resident’s abilities beyond a fixed number of procedures, achieving a tangible improvement in technical skills that will likely result in superior patient safety. To our knowledge, this is the first direct comparison of conventional and proficiency-based simulator training methodologies within the field of otorhinolaryngology, so this is a crucial finding in determining the future utility of surgical simulation.

Submitted for Publication: June 30, 2012; accepted August 16, 2012.

Published Online: October 15, 2012. doi:10.1001/2013.jamaoto.337

Correspondence: Marvin P. Fried, MD, Department of Otorhinolaryngology—Head and Neck Surgery, Montefiore Medical Center, 3400 Bainbridge Ave, Third Floor, Bronx, NY 10463 (mfried@montefiore.org).

Author Contributions: All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Fried, Kaye, Gibber, and Sadoughi. Acquisition of data: Kaye, Gibber, Jackman, Pashkovker, Sadoughi, Schiff, Fraioli, and Jacobs. Analysis and interpretation of data: Fried, Kaye, Pashkovker, and Schiff. Drafting of the manuscript: Kaye, Pashkovker, Schiff, and Jacobs. Critical revision of the manuscript for intellectual content: Fried, Kaye, Gibber, Jackman, Pashkovker, Sadoughi, Schiff, and Fraioli. Statistical analysis: Kaye and Pashkovker. Obtained funding: Fried. Administrative, technical, and material support: Fried, Kaye, Gibber, Jackman, Pashkovker, Sadoughi, Schiff, and Fraioli. Study supervision: Fried, Kaye, Gibber, Jackman, Sadoughi, Schiff, and Jacobs.

Conflict of Interest Disclosures: Dr Fried reports that he is a consultant for Medtronic. Dr Jacobs reports that he is a consultant for GE Navigation and Hemostasis LLC.

Funding/Sponsorship: This study was supported by contract W81XWH-05-1-0577 from the Telemedicine and Advanced Technology Research Center (a division of the US Army Medical Research and Materiel Command) from August 24, 2005, through September 23, 2010.

Previous Presentation: This study was presented at the Combined Otolaryngology Spring Meetings of the Triological Society; April 30, 2011; Chicago, Illinois.

Additional Contributions: We acknowledge the following individuals for their contributions: Seth Lebowitz, MD; Biana Lanson, MD; Seth Lieberman, MD; Joseph Jacob, MD; Yosef Gerson, BS; Bryce Tischkenel, MD; Joshua Silver, MD; Michael Zeltsan, MS; Konrad Schroder, BS; Clarence T. Sasaki, MD; and Douglas A. Ross, MD.

REFERENCES