Hypothesis: Expert levels can be developed for use as training end points for a basic video-trainer skills curriculum, and the levels developed will be suitable for training.

Design: Fifty subjects with minimal prior simulator exposure were enrolled using an institutional review board–approved protocol. As a measure of baseline performance, medical students (n=11) and surgery residents (n=39) completed 3 trials on each of 5 validated video-trainer tasks. Four board-certified surgeons established as laparoscopic experts (with more than 250 basic and more than 50 advanced cases) performed 11 trials on each of the 5 tasks. The mean score was determined and outliers (>2 SDs) were trimmed; the trimmed mean was used as the competency level. Baseline performance of each subject was compared with the competency level for each task.

Setting: All research was performed in a laparoscopic skills training and simulation laboratory.

Participants: Medical students, surgical residents, and board-certified surgeons.

Main Outcome Measures: Expert scores based on completion time and the number of subjects achieving these scores at baseline testing.

Results: For all tasks combined, the competency level was reached by 6% of subjects by the third trial; 73% of these subjects were chief residents, and none were medical students.

Conclusions: These data suggest that the competency level is suitably challenging for novices but is achievable for subjects with more experience. Implementation of this performance criterion may allow trainees to reliably achieve maximal benefit while minimizing unnecessary training.

Arch Surg. 2005;140:80-84
The purpose of this study was to develop and test the suitability of expert levels for a basic video-trainer skills curriculum.

METHODS

All research was performed in the Tulane Center for Minimally Invasive Surgery simulator and training laboratory using a 6-station video-trainer simulator (Karl Storz Endoscopy, Culver City, Calif) (Figure 1). Each module includes a 21 cm × 27.5-cm video monitor (Sony Corporation of America, New York, NY), a Xenon-nova light source, a Telecom SL camera system, a Hopkins II laparoscope, and a Plexiglas box trainer equipped with 1 of 5 previously validated tasks (Karl Storz Endoscopy). The exercises included (in order of increasing difficulty) bean drop, running string, checkerboard, block move, and suture foam (Figure 1), as previously described. Scoring is based on completion time (seconds) using a stopwatch (Fisher Scientific, Hampton, NH). Two research assistants provided all subjects with a brief tutorial concerning the rules of each task and supervised testing but did not give active instruction. Prior to testing, no practice was allowed, except for a single unscored trial on suture foam, so that each subject could become familiar with the Endostitch device (US Surgical Corp, Norwalk, Conn).

Board-certified surgeons (n=4), established as laparoscopic experts (with more than 250 basic and more than 50 advanced cases based on personal case logs), volunteered to develop competency levels. Each expert performed 11 consecutive repetitions on each of the 5 tasks. The mean and SD for all scores were determined using Microsoft Excel (Microsoft Inc, Redmond, Wash). A trimmed mean was obtained by removing outliers (scores greater or less than 2 SDs from the mean) and recalculating. The trimmed mean was rounded to the nearest second and defined as the competency level.

Medical students (n=11, third year) and surgery residents (n=39, postgraduate year 1 [PGY-1] to postgraduate year 5 [PGY-5]) were enrolled in a protocol approved by the Tulane University institutional review board, and informed consent and demographic data were obtained. No student or resident had previous experience with the video-trainer tasks; all residents had previously been tested on the Minimally Invasive Surgical Trainer, Virtual Reality (MIST VR) simulator (Mentice Inc, San Diego, Calif), with less than 2 hours of previous simulator exposure. Each subject completed 3 consecutive repetitions on each of the 5 tasks as a measure of baseline performance. The best score for each subject on each task was compared graphically with the competency levels using Sigma Plot software (SPSS Inc, Chicago, Ill).

Four expert surgeons completed 11 repetitions of each task over a 2-hour period. All were male and right-hand dominant. Expert scores and competency levels are shown...
in Table 1. Only 6% of the raw scores were considered outliers and trimmed. The expert with extensive exposure to video-trainer tasks (expert 1) achieved the best mean score in only 2 of the 5 tasks.

Eleven students and 39 residents completed 3 repetitions of each task over a 1-hour period. Seventy-five percent were male, and 90% were right-hand dominant. Graphs comparing each subject’s best score with the competency levels are shown in Figure 2. For all tasks combined, 6% of subjects reached the competency level. As the task difficulty increased, the number of subjects achieving the levels decreased (Table 2). Seventy-three percent of the successful attempts were made by PGY-5 residents, 13% by PGY-4 residents, 7% by PGY-3 residents, 0% by PGY-2 residents, 7% by PGY-1 residents, and no medical student achieved the competency level.

Two pilot-subject medical students achieved the competency levels on 2 consecutive repetitions for all 5 video-trainer tasks. The mean time to reach competency was 163 minutes, and the mean number of attempts was 90.

Table 1. Expert Performance Scores *

<table>
<thead>
<tr>
<th>Expert</th>
<th>Bean Drop</th>
<th>Running String</th>
<th>Checkerboard</th>
<th>Block Move</th>
<th>Suture Foam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>21.3 ± 3.3</td>
<td>25.7 ± 2.0</td>
<td>69.0 ± 8.4</td>
<td>16.1 ± 1.8</td>
<td>12.6 ± 1.4</td>
</tr>
<tr>
<td>Expert 2</td>
<td>26.6 ± 5.5</td>
<td>32.0 ± 5.3</td>
<td>59.6 ± 8.4</td>
<td>11.9 ± 3.0</td>
<td>21.1 ± 4.9</td>
</tr>
<tr>
<td>Expert 3</td>
<td>23.6 ± 2.6</td>
<td>22.8 ± 2.1</td>
<td>62.2 ± 4.9</td>
<td>17.0 ± 2.5</td>
<td>14.0 ± 3.6</td>
</tr>
<tr>
<td>Expert 4</td>
<td>25.8 ± 5.0</td>
<td>33.7 ± 10.3</td>
<td>95.6 ± 19.0</td>
<td>17.8 ± 2.3</td>
<td>22.6 ± 5.6</td>
</tr>
</tbody>
</table>

*Scores in seconds. Values are mean ± SD from 11 repetitions.

†Competency level defined as the trimmed mean rounded to the nearest second.

Despite the recent surge in popularity of simulator-based training, there is no consensus as to the optimal simulator type or curriculum design. While virtual-reality technology holds great promise, trainees may prefer the video-trainer platform because of issues concerning tactile feedback and imaging fidelity. The Southwestern stations include 3 tasks modified from Rosser et al. so that an assistant is unnecessary, as well as 2 additional tasks that teach depth perception and Endostitch suturing. Improvement in operative performance has been demonstrated after 5 hours of practice using these stations. A subsequent study calculated plateaus in performance and concluded that 30 to 35 repetitions may be a more suitable end point than 5 hours of training. However, calculating performance plateaus for each trainee is cumbersome and not practical for real-time analyses. Moreover, using plateaus may not fully motivate trainees to reach their full potential. A goal-oriented approach may enhance efficiency and maximize learning, although this has not yet been studied.

Performance standards have not yet been developed for most available simulators. The only study reported for a video-trainer platform used the MISTELS tasks, divided trainees into competent and noncompetent groups based on their level of clinical experience, and established a pass/fail criterion based on score distribution. After examining several cutoff scores, the authors identified the best pass/fail standard as having a sensitivity of 82% and a specificity of 82%; thus, 18% of the noncompetent surgeons in the study would have passed, and 18% of competent surgeons would have failed. While this represents one of the first attempts to methodically develop a performance standard, the study’s ultimate goal seems to have been to assess surgeons in practice (to discriminate competent surgeons from noncompetent ones), which is a very high-stakes area that will require extensive validation before being used for credentialing purposes.

Our goal was slightly different; we wanted to develop a performance standard that could be used for training purposes. Our criterion was that the standard would represent a level of mastery achieved by experts for any given skill set but still be attainable by students and residents with practice. With little precedent to rely on, we tried a number of methods and chose those described in this article as optimal. Rather than recruiting a large number of practicing surgeons (who may have had wide variability in their abilities), we identified 4 members of our faculty who not only met our criterion for a minimum number of basic and advanced cases but were also known to be excellent laparoscopists and to have suitable skill sets. To accumulate sufficient data, each expert performed 11 repetitions of 5 tasks. Even though each surgeon was hand-picked, variability between individual surgeon scores for various tasks occurred (Table 1). Such variability was not surprising since an expert often makes up for a relative deficit in one skill set by having a relative surplus of another; for almost every task, a different expert achieved the best score. Despite some individual variability, very few scores needed to be trimmed to produce an overall homogeneous data set. With 206 expert data points, we felt confident that the competency levels obtained were a reasonably accurate representation of mastery for each skill.

As a gauge of suitability, we compared the competency levels with performance data of students and residents (Figure 2). They had no prior video-trainer training and minimal to no prior simulator exposure; testing represented baseline ability. If our competency levels were suitable, very few of the subjects would reach competency at baseline, and this was the case (Table 2). Moreover, the vast majority (86%) who achieved competency were senior residents (PGY-4-5), and no medical student did so, supporting the conclusion that scores were reflective of clinical experience. The only task for which no subject demonstrated baseline competency was the suture foam drill; this was not surprising since none...
had previous training using the Endostitch device. Although using an Endostitch is much simpler than suturing with conventional laparoscopic needle drivers, the device still requires familiarity and practice; suture foam was thus the most difficult task.

When group scores are examined, a trend of improved performance with increased training year is noted. The performance differences seen are likely due to traditional operative experience, thus suggesting construct validity. Unfortunately, these data are an incidental finding and not robust enough to draw this conclusion. Specifically, the power analysis revealed that a much larger sample size would be needed (n=103, α=.80 for P<.05) before the issue of construct validity can be addressed. With continued accrual of subjects, this will be addressed but is not the focus of this article.

The major advantage of using goal-oriented training is uniformity of the final product; all trainees should reach competency. Moreover, subjects can learn at their own speed. This method may efficiently train those with a

Figure 2. Comparison of each subject’s best baseline score with the competency level for all tasks. MS-3 indicates third-year medical students; PGY, postgraduate-year residents.
higher innate ability and avoid unnecessary repetition while ensuring adequate training of those with slower skill development or less innate ability. The ultimate amount of improvement may also be better than the improvement after training using arbitrary duration endpoints, since students and residents function as competitive adult learners and thrive from having a goal to achieve. With a basic skill set already acquired, subjects may then safely enter the clinical arena and care for patients, which undoubtedly will enhance educational opportunities. If proficiency in basic skills is present, time in the operating room can be spent on learning anatomy, pathology, and operative technique instead of skill acquisition.

While this study provides much needed standards for training, additional studies are warranted. Although data from previous studies[10,17,22,24,25] and preliminary work using pilot subjects in our laboratory suggest that most subjects can achieve the competency levels with practice, we have yet to determine whether these levels will be attainable for all trainees. The question then arises whether trainees who cannot reach competency should be allowed to perform procedures that require these skills. High-stakes validation will be required before this question can be answered.

An interesting corollary also exists. Namely, will the use of competency levels discourage exceptionally talented individuals from reaching their full potential? In other words, will using a competency level prematurely terminate a practicing session for someone capable of surpassing an expert level if they continued practicing? Fortunately, with the goal of ensuring a minimum level of mastery, exceptional individuals can safely refine skills further in either the laboratory or the operating room.

In conclusion, the competency standards developed in this study seem suitable for the purpose of performance-based training. The use of such standards should enhance efficiency and more reliably ensure competency. As we continue to enhance surgical education using simulator-based training, careful scrutiny and scientific rigor are needed to optimize training methods and to develop methods for assessment.

### Table 2. Number of Subjects Reaching Competency* During Baseline Testing†

<table>
<thead>
<tr>
<th></th>
<th>Bead Drop</th>
<th>Running String</th>
<th>Checkerboard</th>
<th>Block Move</th>
<th>Suture Foam</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-3 (n = 11)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PGY-1 (n = 19)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PGY-2 (n = 5)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PGY-3 (n = 6)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PGY-4 (n = 4)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PGY-5 (n = 5)</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total (N = 50)</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations: MS-3, third-year medical students; PGY, postgraduate-year residents.

*See Table 1 for competency levels.
†Baseline testing defined as the best score from 3 repetitions.

### REFERENCES