Relationship Between Objective Assessment of Technical Skills and Subjective In-Training Evaluations in Surgical Residents

Liane S Feldman, MD, FACS, Sarah E Hagarty, MD, Gabriela Ghitulescu, MD, Donna Stanbridge, RN, Gerald M Fried, MD, FACS

BACKGROUND: Technical skills of residents have traditionally been evaluated using subjective In-Training Evaluation Reports (ITERs). We have developed the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS), an objective measure of laparoscopic technical ability. The purpose of the study was to assess the concurrent validity of the MISTELS by exploring the relationship between MISTELS score and ITER assessment.

STUDY DESIGN: Fifty surgery residents were assessed on the MISTELS system. Concurrent ITER assessments of technical skill were collected, and the proportion of superior ratings for the year was calculated. Statistical comparisons were performed by ANOVA and chi-square analysis. The Pearson correlation coefficient was used to compare the scores in the MISTELS with the ITER ratings.

RESULTS: The 50 residents received 277 ITERs for the year, of which 103 (37%) were “superior,” 170 (61%) “satisfactory,” 4 (1%) “borderline,” and 0 “unsatisfactory.” The MISTELS score correlated moderately well with the proportion of superior ITER scores ($r = 0.51$, $p < 0.01$). Residents who passed the MISTELS had a higher proportion of superior ITER assessments than those who failed the MISTELS ($p = 0.02$), but residents who performed below their expected level on the MISTELS still received mainly satisfactory ITERs ($82 \pm 18\%$).

CONCLUSIONS: The ITER assessment is poor at identifying residents with below-average technical skills. Residents who perform well in the MISTELS laparoscopic simulator also have better ITER evaluations, providing evidence for the concurrent validity of the MISTELS. Multiple assessment instruments are recommended for assessment of technical competency. (J Am Coll Surg 2004; 198:105–110. © 2004 by the American College of Surgeons)
for time and precision, and an overall score is calculated. Mean scores with 95% confidence intervals have been determined for various levels of training and an overall pass/fail score has been established. To be a useful training and evaluation tool the validity and reliability of the MISTELS must be demonstrated over a series of experiments. Previous work has provided evidence for the construct validity and reliability (test-retest 0.89 and interrater 0.99, unpublished data) of the system. Studies by our group and others suggest that skills learned in a laparoscopic box or virtual reality trainer may be transferable to the operating room environment.

When other scales of the same attribute exist, comparison of results from the two instruments yields information about the concurrent validity of the new instrument. The overall purpose of this study was to assess the concurrent validity of the MISTELS, by examining whether performance in the MISTELS laparoscopic simulator, an objective assessment of technical skill, correlates with the In-Training Evaluation Report (ITER), a widely used subjective assessment of technical skill.

METHODS

Between 1997 and 2000, 116 MISTELS assessments of McGill surgical trainees were identified from a prospectively maintained database. Eliminating medical students, repeated evaluations of the same individual, and entries with missing demographic or scoring data left 60 residents. Complete ITERs for the year the resident was assessed in the MISTELS were identified for 50 residents (35 men, 15 women), who formed the study sample. Residents were grouped into junior (PGY 1–2, n = 20) intermediate (PGY 3–4, n = 21), and senior (PGY 5, n = 9) categories.

MISTELS

The MISTELS program consists of five standardized exercises, performed in a video-trainer, and has been described in detail elsewhere. Briefly, the simulator consists of a trainer box (40 × 30 × 19.5 cm USSC Lap-trainer, United States Surgical Corporation) covered by an opaque membrane. Two 12-mm trocars (USSC Surgiport) are placed according to standard protocol through the membrane at convenient working angles on either side of a 10-mm zero-degree laparoscope (USSC Surgiview) connected to a video monitor. The five tasks range from basic to more advanced laparoscopic skills and include peg transfer, pattern cutting, endoloop placement, intracorporeal knot tying, and extracorporeal knot tying. A 20-minute introductory video demonstrating correct performance of each task is shown to subjects before testing.

The peg transfer involves lifting up six pegs from one pegboard using the left hand, transferring each to the right hand, placing each peg on a second pegboard, then reversing the exercise from right to left. The pattern-cutting task involves cutting a 4-cm diameter, premarked circle from a 10 × 10-cm gauze suspended between alligator clips. For the endoloop task, the subject places and tightens a commercially available pretied slip-knot (USSC Surgitie) on a tubular foam appendage. In the knot-tying tasks, a simple suture is placed through premarked positions in a slit Penrose drain, and the suture is tied using an intracorporeal or extracorporeal knot.

Performance of each task was scored for speed and precision by an observer blinded to ITER assessment. For each task a time score was calculated by subtracting the actual time to complete the task from a preset cut-off time. This gave higher scores to faster performers. If the time to complete the task exceeded the cut-off, the score was recorded as zero. No negative scores were assigned. Precision was factored in with a penalty score unique to each task. The final score was calculated for each exercise by subtracting the penalty score from the performance time score. Normalized scores were calculated by dividing each task score by a constant for the task, and multiplying by 100. This resulted in scores falling in a similar range for each task. Higher scores reflected faster and more accurate performance in each task. The total MISTELS score was calculated by summing the normalized scores for each of the five tasks.

ITER scoring

Surgical residents are assessed after each 2- to 4-month rotation by attending surgeons with standardized In-Training Evaluation Reports (ITERs). ITERs include assessment of three broad categories of competency: knowledge, skills and attitudes, and a global evaluation of competence, with space for comments. Within these categories, 16 specific attributes, ranging from “knowledge of clinical sciences” to “interprofessional relationships” are rated as superior, satisfactory, borderline, unsatisfactory, or could not judge. Question 10 asks the evaluator to rate the resident’s “technical skill” in this manner. A resident receiving a global evaluation of “bor-
derline” or “unsatisfactory” receives no credit for the rotation. The ITERs are kept in the resident’s permanent file in the program director’s office.

The date of the MISTELS evaluation was noted for each resident. ITER evaluations of technical skill for each rotation for the PGY year in which the resident was tested with the MISTELS program were retrieved by an administrative assistant blinded to the purpose of the study. In order to maintain anonymity during data analysis, the residents were coded by number for both ITER grades and MISTELS scores rather than by name. The proportion of “superior” ITER technical skills scores was calculated for each resident.

**RESULTS**

The 50 residents received a total of 277 ITERs for technical skill, of which 103 (37%) were “superior,” 170 (61%) “satisfactory,” and 4 (1%) “borderline.” No resident received an “unsatisfactory” grade. The mean number of ITERs per resident per year was 5.5 ± 1.5 (range 3 to 8) and did not vary significantly by PGY year. Although the trend was for the more senior residents to receive more superior evaluations, this was not statistically significant (Table 1).

To investigate the relationship between resident performances in the MISTELS and ITER, residents were divided into quartiles based on ITER scores (group 1: 0% to 19% superior ITERs; group 2: 20% to 34% superior ITERs; group 3: 35% to 74% superior ITERs; and group 4: 75% to 100% superior ITERs). Residents with poorer ITER grades also performed less well in four of the five MISTELS tasks, and had significantly lower total scores (Table 2).

There was a significant relationship between each MISTELS task and ITER assessment (Table 3). As the score on the MISTELS increased, so did the percentage of superior ITER scores. Because of the known relationship between PGY and MISTELS scores, and because higher-level PGY residents tended to have better ITER grades, a multivariable analysis was performed, includ-

### Table 1. In-Training Evaluation Reports per PGY level

<table>
<thead>
<tr>
<th>Level</th>
<th>n</th>
<th>Number of ITERs/y</th>
<th>Superior ITER (%)</th>
<th>Satisfactory ITER (%)</th>
<th>Borderline ITER (%)</th>
<th>Unsatisfactory ITER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGY 1–2</td>
<td>20</td>
<td>6 ± 1.5</td>
<td>28 ± 28</td>
<td>71 ± 26</td>
<td>2 ± 5</td>
<td>0</td>
</tr>
<tr>
<td>PGY 3–4</td>
<td>21</td>
<td>5.5 ± 1.4</td>
<td>45 ± 29</td>
<td>54 ± 28</td>
<td>1 ± 4</td>
<td>0</td>
</tr>
<tr>
<td>PGY 5</td>
<td>9</td>
<td>4.7 ± 1</td>
<td>47 ± 36</td>
<td>51 ± 36</td>
<td>2 ± 7</td>
<td>0</td>
</tr>
</tbody>
</table>

p 0.09 0.2 0.18 0.85 1

Data expressed as mean ± SD.

ITERs, In-Training Evaluation Reports.

### Table 2. McGill Inanimate System for Training and Evaluation of Laparoscopic Skills Scores Stratified by In-Training Evaluation Reports Quartile

<table>
<thead>
<tr>
<th>ITER quartile</th>
<th>n</th>
<th>Peg Cutting</th>
<th>Endoloop</th>
<th>IC knot</th>
<th>EC knot</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>32 ± 30</td>
<td>45 ± 25</td>
<td>38 ± 33</td>
<td>39 ± 31</td>
<td>187 ± 125</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>33 ± 27</td>
<td>55 ± 25</td>
<td>63 ± 16</td>
<td>47 ± 28</td>
<td>242 ± 68</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>52 ± 20</td>
<td>66 ± 15</td>
<td>61 ± 30</td>
<td>65 ± 28</td>
<td>296 ± 69</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>62 ± 19</td>
<td>77 ± 8</td>
<td>73 ± 16</td>
<td>84 ± 15</td>
<td>346 ± 43</td>
</tr>
</tbody>
</table>

p 0.016 0.003 0.014 0.002 0.07 0.0004

Data expressed as mean ± SD.

Group 1: 0% to 19% superior; group 2: 20% to 34% superior; group 3: 35% to 74% superior; and group 4: 75% to 100% superior.

EC, extracorporeal; IC, intracorporeal; ITER, In-Training Evaluation Report.

### Statistical analysis

Data are expressed as mean ± standard deviation, unless otherwise stated. Statistical comparisons were performed by one-way analysis of variance (ANOVA) and chi-square analysis. The correlation between the MISTELS total score and the proportion of superior ITER technical skills scores was assessed using the Pearson coefficient. Because of the effect of PGY level on MISTELS score, a multivariable analysis was performed to assess the relationship between MISTELS and ITER scores, and p < 0.05 was considered statistically significant.
Table 3. Correlation Between McGill Inanimate System for Training and Evaluation of Laparoscopic Skills Scores and In-Training Evaluation Reports Scores

<table>
<thead>
<tr>
<th>MISTELS task</th>
<th>ITER Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peg transfer</td>
<td>0.45 (0.0001)</td>
</tr>
<tr>
<td>Pattern cutting</td>
<td>0.45 (0.0001)</td>
</tr>
<tr>
<td>Endoloop placement</td>
<td>0.35 (0.01)</td>
</tr>
<tr>
<td>Intracorporeal knot</td>
<td>0.45 (0.0001)</td>
</tr>
<tr>
<td>Extracorporeal knot</td>
<td>0.29 (0.04)</td>
</tr>
<tr>
<td>Total score</td>
<td>0.51 (0.0001)</td>
</tr>
</tbody>
</table>

Values are Pearson correlation coefficients; p values are in parentheses; and the percent of superior ITERs was used in analysis.

ITERs, In-training Evaluation Reports; MISTELS, McGill Inanimate System for Training and Evaluation of Laparoscopic Skills.

Table 4. Residents Who Passed the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills Also Had Higher In-training Evaluation Reports Scores

<table>
<thead>
<tr>
<th>MISTELS outcome</th>
<th>n</th>
<th>Superior ITER (%)</th>
<th>Satisfactory ITER (%)</th>
<th>Borderline ITER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail</td>
<td>26</td>
<td>29 ± 25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pass</td>
<td>24</td>
<td>49 ± 32</td>
<td></td>
<td>0.02</td>
</tr>
</tbody>
</table>


DISCUSSION

We found a relationship between objective performance on the MISTELS laparoscopic trainer and subjective assessment of technical skill in surgical residents. Scores in four of five MISTELS tasks (all except extracorporeal knot tying) and MISTELS total score were found to positively correlate with ITER assessment. Residents who passed the MISTELS had higher ITER grades than residents who failed MISTELS. Residents performing below expected levels on the MISTELS also performed less well on the ITERs. These results provide evidence for the concurrent validity of the MISTELS in evaluating technical skills.

The MISTELS total score correlated moderately well with the percent of superior ITERs (r = 0.51), falling within the range of correlation expected when comparing two measures of the same attribute (0.4 to 0.8).** So the MISTELS total score accounts for about 25% of the variability in ITER assessment. Which other factors might affect ITER assessment? The ITER assessment of technical skill is a global evaluation and likely reflects both decision making and technical skill in the operating room; MISTELS reflects psychomotor performance alone. The discrepancy between MISTELS and ITER scoring is in line with the widely quoted estimation that “. . . about 75% of the important events in an operation are related to making decisions, and about 25% to dexterity.”

Although widely used, the validity of the ITER as a measurement tool has been repeatedly challenged and problems with the ITER system were apparent in this study. First, ITER scores were essentially all superior or satisfactory. The mean percentage of superior evaluations per year was 38.5 ± 30%. It was distinctly unusual for a resident to receive a “borderline” assessment, probably because this grade leads to automatic review by the program director, and may result in having to repeat the rotation. There were no unsatisfactory evaluations. Differences between residents were apparent only when the proportion of “superior” evaluations was calculated for the year, as would be done for a resident’s year-end summary assessment by the program director. So residents scoring in the lowest group on the MISTELS still obtained mainly “satisfactory” ITERs throughout the year. If the resident and the training program are relying solely on ITERs, residents struggling with their technical skills are unlikely to be readily identified and the chance for

Table 5. Residents Performing below Expected Level on the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills Also Had Poorer In-Training Evaluation Reports Assessments

<table>
<thead>
<tr>
<th>MISTELS total score</th>
<th>n</th>
<th>Superior ITER (%)</th>
<th>Satisfactory ITER (%)</th>
<th>Borderline ITER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; Expected for level</td>
<td>15</td>
<td>16 ± 20</td>
<td>82 ± 18</td>
<td>2 ± 5</td>
</tr>
<tr>
<td>= Expected for level</td>
<td>12</td>
<td>38 ± 27</td>
<td>60 ± 28</td>
<td>2 ± 6</td>
</tr>
<tr>
<td>&gt; Expected for level</td>
<td>23</td>
<td>53 ± 30</td>
<td>46 ± 29</td>
<td>1 ± 4</td>
</tr>
</tbody>
</table>

Data expressed as mean ± SD.

timely remedial training may be missed. This failure of evaluators to use the entire rating scale has been termed the “central tendency error.”

In our program, the significant penalties resulting from rating a resident below average negatively impact on the ITER’s utility as a feedback mechanism.

Other rater errors decrease the validity of ITERs as measurements of technical skill. The “halo effect” refers to the fact that performance in one area can cloud judgment of other aspects of performance. For example, Warf and colleagues found that senior residents were more likely than interns to be judged “competent” in an Objective Structured Clinical Examination, even though there was no difference in the specific skill being tested. In order to address the possible impact of PGY level on ITER score, we stratified the MISTELS scores according to previously defined “expected” levels for PGY level, and included PGY as a variable in a model predicting ITER. We did not find that PGY independently predicted ITER when the MISTELS score was taken into account.

Despite the drawbacks of the ITER, it remains in widespread use, so was chosen to represent a measure of technical skill against which to judge the MISTELS. We have previously assessed concurrent validity of the MISTELS by comparing performance in the MISTELS with performance of analogous tasks in a pig model, demonstrating good correlation in five of seven tasks.

Ideally, the concurrent validity of the MISTELS would be judged using intraoperative objective assessment scales, such as those described by Wanzel and associates. These studies are being planned.

Several weaknesses are identifiable in this study. First of all, residents were judged on their skill after only one MISTELS attempt. This would particularly penalize the junior residents, who improve quickly after their initial unfamiliarity with the system is overcome. In addition, the study sample may have had practice opportunities on the simulators that were not recorded in the database. For example, we found a much higher than anticipated number of residents performing above their expected level, suggesting the group had exposure to MISTELS in the past. Finally, MISTELS was designed to test laparoscopic skills; the ITER is a global assessment of all technical skill. We have not assessed how MISTELS correlates with other objective measures of nonlaparoscopic surgical skill.

Although the ITER is far from an ideal instrument for measuring technical skill, it is widely used, and the relationship between MISTELS and ITER scores supports the notion that some component of technical ability is measured by the ITER. A reconfigured ITER, including a more detailed rating scale that does not harshly penalize the weakest residents may be more informative for trainees. But the era of the ITER as the sole arbiter of technical competence is probably ending. Objective measures, like MISTELS for laparoscopic skills, provide immediate feedback, an assessment of where the trainee stands in relation to his or her peers, and a definition of competence derived in a defensible manner. The use of multiple assessment tools to define competence and aid in training is suggested.

**Author Contributions**

Study conception and design: Feldman, Fried

Acquisition of data: Stanbridge, Hagarty, Ghitulescu

Analysis and interpretation of data: Feldman, Hagarty

Drafting of manuscript: Feldman, Hagarty

Critical revision: Fried

Statistical expertise: Feldman

Obtaining funding: Fried

Supervision: Fried, Feldman

**REFERENCES**


