World J Surg (2008) 32:156–160
DOI 10.1007/s00268-007-9143-y

World Journal of Surgery

Objective Assessment of Technical Performance

Gerald M. Fried · Liane S. Feldman

Published online: 12 June 2007

Abstract Technical skills are essential to the practice of surgery. They can be taught in the operating room and in the surgical skills laboratory. The use of simulators allows the development of reproducible curricula with measurement of performance using objective metrics. The goal of those designing metrics for the simulation laboratory should be to establish measures that are consistent with those of high quality surgery in the operating room. Once these metrics have been shown to be reliable, valid, practical to use, and meaningful to the learner, they can form the basis of a learning program based on the acquisition of proficiency. Performance in the skills laboratory should ultimately be predictive of performance in the clinical setting.

Can we measure (technical) surgical skills reliably and objectively? How should we use this information? Surgical educators around the world are contemplating these issues. The answers to these questions have the potential to play an important role in the evolution of surgical training from an apprenticeship-based system to one based on structured curricula with clear, objective goals for technical proficiency.

Operations can be deconstructed, and skills can be identified to be simulated. Surgical technical skills can then be taught in the skills laboratory with the use of simulation materials and task trainers. An advantage of education using simulation is the potential to measure performance and establish objective metrics to be used for formative and summative assessments. This enables the development of proficiency-based curricula rather than time-based educational programs.

Objective measures of performance require the development of metrics and evidence that these measures are practical to use, meaningful to the learner and the teacher, reliable, and valid. Ideally, measurements taken in the skills laboratory should reflect values of quality performance in the clinical setting and be predictive of performance in the operating room.

Measurement of technical skills in the clinical setting

Technical skills are constantly evaluated during residency training. The goal of these assessments is to ensure the appropriate development of surgical skills and provide feedback to the learner. These assessments generally are in the form of periodic in-training evaluation reports (ITERs) that are used to assess all areas of competence of the trainee. Unfortunately, when addressing technical skills, these evaluations are generally subjective, vague, unreliable, lack validity, and rarely provide meaningful specific suggestions to remediate areas of deficiency [1, 2]. They are based on direct observation by one or more supervising attending surgeons, and the feedback is distant in time from the events on which the assessments are based. During training, the technical skills being evaluated are subject to the vagaries of the clinical material to which the resident is exposed and can vary greatly among residents in any given peer group.

ITERs have poor interrater reliability and modest or poor validity. They are subject to halo effects, where well-liked residents, or those who are intelligent and responsible, are evaluated better on their technical skills than res-

G. M. Fried (✉) · L. S. Feldman
Department of Surgery, Steinberg-Bernstein Centre for Minimally Invasive Surgery & Innovation, McGill University Health Centre, 1650 Cedar Avenue, #L9.309, Montreal, Quebec, Canada H3G 1A4
e-mail: Gerald.fried@mcgill.ca

idents who function less well in nontechnical domains. These ITERs also suffer from failure to use the entire scale. Almost everyone is rated as average or above average.

Although procedural logs are used as a surrogate for technical competence, the number of cases does not equate to competence, and the validity of this measure has not been established. A few efforts have been made to establish objective measures of technical skill in the operating room, but these measures have not been widely adapted yet, in part because of the usual need to have an expert observer present to perform the evaluation [3, 4]. Intraoperative assessments can also be done by videotape review, although this method lacks information about verbal cues and the quality of assistance [5].

The use of simulation allows the development of a reproducible technical challenge and assessment of performance using well designed metrics that withstand scrutiny for their reliability, validity, ease of use, and exportability. Objective assessment of technical skill through simulation is explored in detail below. Fundamental to assessment is the development of credible metrics.

Development of metrics

There are several options available to measure performance. Generally, they can be categorized into efficiency and quality.

Efficiency

When evaluating performances of experts and novices, the most obvious difference is in efficiency. Experts work more quickly and have more purposeful movements. These differences can be readily measured objectively using time to complete the task [6] and motion analysis to track movement of the surgeon using systems such as the ICSAD [7, 8] or tips of the instruments (distance or angles) using virtual reality-based laparoscopic simulators. These metrics can be calculated automatically in physical or virtual reality simulators. By their nature, these measures of efficiency are objective, highly reproducible, and independent of rater interpretations.

Efficiency alone can be misleading and should not in itself be the value sought by the learner. However, when the learner knows that penalties will be assigned for errors, speed must be sacrificed to avoid errors, and efficiency in this context is a valuable metric [6].

Quality

It is mandatory that some measure of quality be included in any assessment. This usually includes some way to track errors. Errors may be assessed in terms of their number, frequency, and degree of importance.

End-product analysis provides important information about the quality of performance. For example, when testing suturing skill, time to complete the task is immaterial if the stitch does not approximate the tissue properly or if the knot is poorly constructed. The completed stitch can be inspected, and the quality of the result can be readily evaluated. In this example, the accuracy of stitch placement may be measured as the difference (in millimeters) between where the needle is passed through the tissue and the location of a target. This generates a continuous variable and is readily measured with low interrater variability. For the same suturing task, an incomplete knot, an insecure knot, or a knot that does not approximate the tissues in question would be an error. These errors must be assessed in a standardized, objective way to minimize the judgment required of the evaluator. Knot security can be easily evaluated, and calipers can be used to measure such factors as gaps in the tissue. An incomplete knot or one that falls apart with tension could be considered a ‘‘critical error’’ and would result in failure no matter how quickly the task were completed.

If a goal of simulation training is to develop good habits, the habits can be assessed using rating scales or checklists. Properly holding instruments, smooth rotation of the wrist as the needle passes through tissues, the sequence of movements while performing tasks, among others, can be assessed. This is more difficult to do automatically and requires well trained evaluators to be able to do these assessments with reasonable interrater reliability. These assessments are valuable for formative evaluation; their role in summative evaluation may be less valuable as they are more difficult to standardize in the operating room.

The Objective Structured Assessment of Technical Skill (OSATS) [9] was developed to assess surgical skill in a standardized and objective way in the skills center. This excellent program has been well validated and shown to be an effective method for assessment, though requiring significant human resources. When studying OSATS, it was evident that global rating scales are more effective than checklists for rating surgical skill.

Because the use of OSATS makes great demands on the time of the faculty, Datta and colleagues [10] developed a more efficient method of assessing technical skills using a combination of end-product evaluation, a ‘‘snapshot’’ assessment of task performance based on an edited 2-minute video recording scored with OSATS and measurement of the surgical efficiency score through a combination of final product quality and hand-motion analysis. Vassiliou et al. [11] also found that global rating scales provided better assessment of surgical skill when evaluating laparoscopic surgeons in the operating room.

Another measure of quality when evaluating surgical skills is the smoothness of the movements of the surgeon while doing a task. Jerky motions can cause injury to adjacent tissues and may reflect lack of skill. Smoothness data can be generated by motion analysis, tracking change, using measurements of velocity, and vectors [12].

Matching simulator and clinical metrics

One of the problems in developing and evaluating simulation metrics is the lack of a gold standard by which to measure technical skill in the operating room. The goal of skills training is to enhance those skills that are important determinants of the quality of surgical performance. Although most would agree that efficiency, precision, and avoidance of errors are qualities that reflect technical skill in the operating room, few have attempted to develop reliable and valid measurements of these attributes. A few efforts have been made to make these measurements and ensure that they meet critical standards of reliability and validity [11, 13]. When these measures are available, they provide a useful means to assess the quality of the simulation metrics.

Measurements made in the skills laboratory should be readily available to learners to provide immediate feedback and allow evaluation of their performance in relation to a standard (e.g., proficiency level, performance of their peers). They should be able to use the information provided by these evaluations to modify their performance and observe their progress. If the evaluation is to provide useful feedback, it must be meaningful to the learner. The use of objective metrics avoids the personal aspects of the evaluation. Metrics that are consistent with values in the clinical setting are especially motivating. It is easy for the student to understand feedback that says, ''Your knot is not square and it slips under tension, so you get a penalty of x points.'' The student can use this feedback and ensure that the next knot is square and tight. In contrast, although feedback about motion analysis is reliable and valid, it may be more difficult to use it to guide behavioral change on the part of the student. How does the learner integrate information that his or her path length measured 6800 cm during a suturing drill?

The most useful metrics for the learner would express performance in the skills laboratory and the operating room in parallel terms. Furthermore, practice using the simulator would result in improved operative skill, and performance in the skills laboratory would be predictive of the performance level observed in the clinical environment.

Evaluation of metrics

Whether developed for the simulator or operating room, measurements of technical skills should meet certain standards. They should be reliable, valid, practical, generalizable, and useful for feedback purposes.

Reliability

Reliability refers to the consistency of the measures and is usually evaluated in three domains: interrater reliability, test-retest reliability, and internal consistency.

Objective assessments should have high interrater reliability. This refers to the consistency of evaluations among evaluators. Interrater reliability is assessed by calculating the intraclass correlation coefficient (ICC). This value ranges from 0 to 1.0, with coefficients >0.80 required for high-stakes evaluations.

Test-retest reliability is also assessed by the ICC and is used to assess the consistency of the evaluation for a given individual between tests. This assumes that the individual has reached a plateau or steady state, and the primary determinant of variance between assessments is the instrument used for assessment. In fact, an individual's performance may vary for many other reasons, such as motivation, fatigue, or distractions.

Because most assessment measures rate performance using more than a single score or domain, a good measure of performance shows internal consistency among the test items used for assessment. The assumption here is that a skilled surgeon will rate highly (although not identically) when assessed in different domains. The measure of internal consistency is Cronbach's α, which is also measured between 0 and 1.0.

Computer-generated results tend to show excellent interrater reliability because there is no human involved with the measurement, and values are generated electronically. Test-retest reliability for computer-based or virtual reality systems may be affected by the lack of familiarity of the subject with the interface. It is important to allow the subject time to become familiarized with the interface before performing the assessment to maximize reliability in this domain.

Human assessment of technical skill may be influenced by the training and expertise of those doing the evaluation as well as the degree of objectivity built into the assessment. With properly trained raters and the use of objective measures of performance, excellent reliability can be achieved [14].

Validity

For a test to be useful, it must be shown to be valid. Because the test is being used to make an assumption about performance in the real world, validity refers to how well the test scores represent the performance they are designed to assess. Another way to think of validation is as a process

of hypothesis testing, ''whereby we determine the degree of confidence we can place on inferences we make about people based on their scores'' [15]. Validation studies are generally performed in four spheres [16]: predictive validity, concurrent validity, content validity, and construct validity.

Predictive and concurrent validity are classified as criterion-oriented validation procedures. If the test score is designed to evaluate technical skill in surgery, the correlation between the test score on the simulator and another measure of technical surgical skill (e.g., performance in the live animal operating room or in the clinical setting) can be used to assess criterion validity. If these evaluations are done closely in time, they are referred to as testing concurrent validity; if the simulator test is done before the clinical test, it is predictive validity that is being assessed. Examples of studies to assess criterion validity for the Fundamentals of Laparoscopic Surgery manual skills test can be seen in the articles by Feldman et al. [17] and Fried et al. [18]. Using a multivariate model, McCluney et al. [19] showed that simulator assessments were highly predictive of operative performance, independent of the level of the surgeon's experience. Using receiver operating characteristics (ROC) methodology, they showed that proficiency levels could be established for performance in the simulator that were sensitive and specific for predicting ''competent'' skill in the operating room.

Content validity is established by showing that the test items are a good sample of the skills the evaluator is interested in assessing. It is measured in response to the question, ''Does the test (simulator) evaluate the appropriate (specific) content and breadth of content?''

Validity is generally demonstrated through a series of studies each providing evidence toward establishing the validity of the test. Although a single test may separate experts from novices, a good evaluation provides feedback on the breadth as well as the depth of skill. The development of the MISTELS program to assess laparoscopic skills provides an example of this process. To determine the various modules of the program, experienced surgeons were asked to review videotapes of operations and list the skills they thought were fundamental to laparoscopic surgery. Another group of experts reviewed the MISTELS modules and were asked to complete a global rating scale to determine which of the fundamental areas of minimally invasive surgical skills were represented in the MISTELS program and how well each module and the metrics developed for that module represented that skill [20, 21].

Face validity is a form of content validity and refers to the subjective evaluation of the metrics by experts in the field. It is usually evaluated by a questionnaire that asks, ''On the face of it, do the metrics seem credible measures of the construct in question?'' Although face validity is more subjective than other measures of validity, it is important to gain ''buy in'' from the user.

When there is no good ''gold standard'' to use as a criterion against which the test can be compared, other methods are required to demonstrate the validity of the test instrument. ''Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured'' [16]. Construct is defined as a theoretical construction about the nature of human behavior. It is not directly observable but must be inferred from its observable effect on human behavior. Thus, construct validity refers to the extent to which inferences from a test's score accurately reflect the construct that the test is claiming to measure. One way to assess construct validity is to demonstrate that the test can detect differences between groups with known differences. Because technical skill usually increases with experience, scores in the simulator should follow the same pattern. Groups may be divided based on their level of experience (e.g., novices and then intermediate and experienced surgeons), and performance in the simulator can be compared among these groups. For example, Datta and colleagues developed a standardized test of technical skills for general surgical trainees. With this program, they evaluated open, laparoscopic, and flexible endoscopy skills using a variety of assessment strategies, including global rating scales based on the OSATS model as well as computer-generated scores. [22]. By comparing groups with known differences (basic surgical trainees, junior specialist trainees, senior specialist trainees), they were able to show good construct validity for the summary metrics used in this program.

Many reported studies have shown significant differences in performance between small groups with large expected differences in skills (e.g., practicing surgeons versus medical students). Large studies are required to assess whether the simulator metrics can discriminate finer differences, such as those expected from year to year in residency training [6, 13, 23, 24].

Once groups have been defined based on some accepted construct (e.g., level of experience), the test can be assessed for sensitivity, specificity, and positive and negative predictive values when using any specific passing score to determine ''competence.'' Construction of an ROC curve involves plotting the sensitivity versus 1 - specificity of the test for any given passing score and helps the evaluator determine the best passing score for the purpose of the test [25].

One more measure of the validity of the test assesses the generalizability of the test and its metrics. This is used to evaluate the extent to which the results of a research study can be generalized to individuals and situations beyond those involved in the study. For widespread adoption of an

assessment, it needs to be shown to be generalizable and effective when testing diverse groups [20, 21].

Conclusions

Several methods are available to us to evaluate surgical technical skills using objective techniques. These methods can be applied in the operating room, animal or cadaver laboratory, or the simulation laboratory. Metrics used to evaluate technical skills can be assessed in a methodical way to ensure that these measures are reliable, valid, practical, and meaningful and can be generalized.

References

1. Moorthy K, Munz Y, Sarker SK, et al. (2003) Objective assessment of technical skills in surgery. BMJ 327:1032–1037
2. Sidhu RS, Grober ED, Musselman LJ, et al. (2004) Assessing competency in surgery: where to begin? Surgery 135:6–20
3. Scott DJ, Bergen PC, Rege RV, et al. (2000) Laparoscopic training on bench models: better and more cost effective than operating room experience? J Am Coll Surg 191:272–283
4. Vassiliou MC, Feldman LS, Andrew CG, et al. (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190:107–113
5. Scott DJ, Rege RV, Bergen PC, et al. (2000) Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech A 10:183–190
6. Derossis AM, Fried GM, Abrahamowicz M, et al. (1998) Development of a model for training and evaluation of laparoscopic skill. Am J Surg 175:482–487
7. Bann SD, Khan MS, Darzi A (2003) Measurement of surgical dexterity using motion analysis of simple bench skills. World J Surg 27:390–394
8. Brydges R, Classen R, Larmer J, et al. (2006) Computer-assisted assessment of one-handed knot tying skills performed within various contexts: a construct validity study. Am J Surg 192:109–113
9. Martin JA, Regehr G, Reznick R, et al. (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84:273–278
10. Datta V, Bann S, Mandalia M, et al. (2006) The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. Am J Surg 192:372–378
11. Vassiliou MC, Feldman LS, Andrew CG, et al. (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190:107–113
12. Van Sickle KR, McClusky DA III, Gallagher AG, et al. (2005) Construct validation of the ProMIS simulator using a novel laparoscopic suturing task. Surg Endosc 19:1227–1231
13. Fried GM, Feldman LS, Vassiliou MC, et al. (2004) Proving the value of simulation in laparoscopic surgery. Ann Surg 240:518–528
14. Vassiliou MC, Ghitulescu GA, Feldman LS, et al. (2006) The MISTELS program to measure technical skill in laparoscopic surgery: evidence for reliability. Surg Endosc 20:744–747
15. Streiner DL, Norman GR (1995) Health Measurement Scales: A Practical Guide to Their Development and Use. 2nd edition. Oxford, Oxford University Press
16. Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52:281–302
17. Feldman LS, Hagarty SE, Ghitulescu G, et al. (2004) Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents J Am Coll Surg 198:105–110
18. Fried GM, Derossis AM, Bothwell J, et al. (1999) Comparison of laparoscopic performance in vivo with performance measured in a laparoscopic simulator. Surg Endosc 13:1077–1081
19. McCluney AL, Vassiliou MC, Stanbridge DD, et al. (2007, in press) FLS Simulator performance predicts intraoperative laparoscopic skill. Surg Endosc
20. Swanstrom LL, Fried GM, Hoffman KI, et al. (2006) Beta test results of a new system assessing competence in laparoscopic surgery. J Am Coll Surg 202:62–69
21. Peters JH, Fried GM, Swanstrom LL, et al. (2004) Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 135:21–27
22. Datta V, Bann S, Aggarwal R, et al. (2006) Technical skills examination for general surgical trainees. Br J Surg 93:1139–1146
23. Derossis AM, Bothwell J, Sigman HH, et al. (1998) The effect of practice on performance in a laparoscopic simulator. Surg Endosc 12:1117–1120
24. Derossis AM, Antoniuk M, Fried GM (1999) Evaluation of laparoscopic skills: a 2-year follow-up during residency training. Can J Surg 42:293–296
25. Fraser SA, Klassen DR, Feldman LS, et al. (2003) Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. Surg Endosc 17:964–967