Testing Technical Skill via an Innovative "Bench Station" Examination

Richard Reznick, MD, Glenn Regehr, PhD, Helen MacRae, MD, Jenepher Martin, MD, Wendy McCulloch, RN, *Toronto, Ontario, Melbourne, Australia*

BACKGROUND: A new approach to testing operative technical skills, the Objective Structured Assessment of Technical Skill (OSATS), formally assesses discrete segments of surgical tasks using bench model simulations. This study examines the interstation reliability and construct validity of a large-scale administration of the OSATS.

METHODS: A 2-hour, eight-station OSATS was administered to 48 general surgery residents. Residents were assessed at each station by one of 48 surgeons who evaluated the resident using two methods of scoring: task-specific checklists and global rating scales.

RESULTS: Interstation reliability was 0.78 for the checklist score, and 0.85 for the global score. Analysis of variance revealed a significant effect of training for both the checklist score, $F(3,44) = 20.08$, $P < 0.001$, and the global score, $F(3,44) = 24.63$, $P < 0.001$.

CONCLUSIONS: The OSATS demonstrates high reliability and construct validity, suggesting that we can effectively measure residents' technical ability outside the operating room using bench model simulations. *Am J Surg.* 1996;172:226–230.

If the public were surveyed as to what qualities are important in a surgeon, technical skill would undoubtedly be near the top of the list. Yet, of all the qualities important to the development of a surgeon measured in our training programs, the formal assessment of technical skill is the weakest. At present, surgical training programs and certification bodies do an excellent job of assessing cognitive knowledge and a good job of assessing surgical judgment. There has been a heightened awareness of the need to assess the clinical skills of data gathering, data interpretation, and patient doctor communication, concepts that can be adequately assessed using existing examination formats such as the Objective Structured Clinical Examination (OSCE).[1] Professional qualities such as honesty, maturity, and diligence are primarily assessed through in-training reports, which usually represent an amalgam of preceptors informal opinions. Competence in the technical domain is, to be sure, taken seriously by all surgical training programs. However, like the domain of professional qualities, the assessment of technical skill has, for the most part, been done informally and in a nonstandardized fashion.

The development of a standardized test of technical skill would serve many purposes. It has the potential to set appropriate standards and levels of expectation for our trainees. A standardized test would be invaluable as a feedback tool for the residents. Residents with deficits could be identified early and remedial programs developed. Such a test could be used in promotion decisions, and would have the potential to validate decisions that at present are being made by the unratified opinions of preceptors. A test of technical skill, if reliable and valid, could allow for interinstitutional comparisons and ultimately could be used as a tool in certification and recertification.

Several years ago, the Surgical Education Research Group at the University of Toronto started a program of research with two goals. The first was to evaluate the efficacy of teaching some aspects of technical skill development outside the operating room in a bench setting. The second, which we viewed as a precondition for achieving our first goal, was the development of a reliable and valid assessment tool for technical skill. Borrowing from the success of the OSCE in the domain of clinical skills, we developed a new examination, the Objective Structured Assessment of Technical Skill (OSATS). This examination is a multistation performance based assessment of technical skill developed by Martin and colleagues.[2] The initial work on this examination was aimed at evaluating the reliability of the examination, comparing a live animal platform to a bench model platform, and assessing two evaluation tools, a task-specific checklist and a global rating approach. Martin et al[2] reported on the results of 20 surgical residents who took the OSATS in two parallel versions: a six-station live animal model examination and a bench model examination where the same six tasks that were tested in the animal environment were tested using bench model simulations. The reliability estimates for both examination platforms were in the moderate to high range (.66 and .74) when a global rating approach was used to scoring, and were mixed (.61 and .33) when a task-specific checklist approach to scoring was used.

There were no significant differences in the psychometric properties of the live animal version compared with the bench model version. Both approaches to scoring, in both platforms of examination, were able to show increasing lev-

From the Department of Surgery (RR, GR, HMac, WMc), University of Toronto, Toronto, Ontario, Canada, and the Department of Surgery (JM), University of Melbourne, Melbourne, Australia.

This work was supported by the physicians of Ontario through a grant given by the P.S.I. foundation.

Requests for reprints should be addressed to Richard K. Reznick, MD, The Toronto Hospital, University of Toronto, EN 9-237, 200 Elizabeth Street, Toronto, Ontario, Canada M5G 2C4.

Manuscript submitted August 13, 1996 and accepted in revised form October 2, 1996.

0002-9610/97/$17.00
PII S0002-9610(97)00023-8

els of competence with increasing years of experience as a surgical resident, ie, construct validity.

The results of this initial experiment encouraged us to proceed with full-scale testing of our general surgical residents on an annual basis. In so doing, a decision was made to rely solely on bench model simulations. This decision was made for a variety of reasons, but was founded in the demonstration of psychometric equivalence of the two formats.

This report presents data from the first large-scale administration of the OSATS in 1995, during which 48 general surgical residents were tested. The specific questions that were addressed in this study were the reliability of an eight-station, 2-hour examination; and the extent to which the examination was construct valid, that is, its ability to differentiate residents at different levels of training.

METHODS

Examination Format

The OSATS is a performance-based examination designed to assess the technical skills of surgical trainees. The examination is consistent with the format of the typical OSCE in which examinees perform a series of clinical tasks at each of several time-limited stations. Stations involve bench model simulations of operative procedures appropriate to general surgery. Eight 15-minute stations were used for the examination, including excision of a skin lesion; insertion of a T-tube, abdominal wall closure; handsewn bowel anastomosis; stapled bowel anastomosis; control of IVC hemorrhage; pyloroplasty; and tracheostomy.

Performance at each station was marked by a qualified surgeon with particular expertise in the procedure being simulated at the station. The examiners marked performances using two evaluation tools. The first evaluation tool was an operation-specific checklist. The operation-specific checklist identifies separate actions felt by a panel of surgeons to be necessary in performing the operative task effectively. Each checklist was composed of 20 to 40 items that are relevant to a specific operation. Examiners marked candidates by indicating the items or actions that the candidate performs competently during the operative task. A resident's score for a given station was the proportion of items checked by the examiner as done correctly at the completion of the station. Since each checklist is operation specific, a different checklist is used for each of the eight stations. A sample of a task-specific checklist may be seen in **Figure 1**.

The second evaluation tool used was a global rating scale. This global rating scale consists of seven dimensions, each related to some aspect of operative performance. Each dimension was graded on a 5-point scale with the middle and extreme points anchored by explicit descriptors. Each 5-point item was scored from 0 (poor performance) to 4 (good performance). A resident's score for a given station was determined by summing the marks on the seven dimensions and dividing by 28 to obtain a percentage score. Items on the global rating scale were developed to be operation independent, so the same rating scale is used at each of the eight stations. Previous research has shown that this global rating scale is a reliable and valid assessment of technical skill both in the operating room[3] and in the simulated operating environment.[2] A copy of the global rating scale may be seen in **Figure 2**.

Three tracks of the eight-station examination were run simultaneously, for a total of 24 concurrent stations. The examination was run twice on the same day using different examiners in the morning and afternoon.

Participants

Forty-eight general surgery residents from the University of Toronto participated as examinees. Length of time in the residency program ranged from 1 to 6 years. Administration of the examination required 48 qualified surgeons to act as examiners, each required for approximately 2 hours. In addition, 8 support staff were required for the full day of the examination to reset models, and 21 nurses acted as operative assistants (1 at each of seven stations for each of three tracks) for the whole day.

Statistical Analysis

Interstation reliability of the global rating scale and the checklist were calculated separately using Cronbach's coefficient alpha. Construct validity of the measures was assessed using two one-way analysis of variance (ANOVA) models with training level as the independent measure and each measure as the dependent variable. For the ANOVAs, an alpha level of 0.05 was used.

RESULTS

The interstation reliability was 0.843 for the global rating scale and 0.781 for the checklist. The means of the global rating scale and the checklist at each level of clinical training are presented in **Figure 3**. The ANOVA on global ratings revealed a significant effect of training level, $F(3,44) = 24.63$, with this variable accounting for 62.7% of the variance in global scores. Further, post-hoc analyses revealed that each level of training showed a significant improvement in global rating scores. Similarly, the ANOVA on checklist scores revealed a significant effect of training level, $F(3,44) = 20.08$, with this variable accounting for 57.8% of the variance in checklist scores. Post-hoc analysis on the checklist scores revealed significant differences between all levels of training, with the exception that there was no significant difference between residents in the PGY4 category compared with residents in the PGY5/6 category.

COMMENTS

To be sure, surgeons responsible for training the surgeons of the future take that job very seriously. There is an implicit obligation to assure that graduates of a training program have satisfied all the educational objectives articulated by that program. For surgical specialties, technical competence is a central construct that lies at the heart of public expectations. Despite this, most training programs rely on in-training reports for the documentation of technical skills. These reports, while undoubtedly the product of serious thought and evaluation, are often vague and imprecise. Terms such as "a good pair of hands" and "progressing nicely" and "a bit green in the OR but to be expected of a junior trainee" are familiar to us all. Despite their familiarity, they are not standardized, are prone to misinterpretation, and cannot adequately serve as the basis for promotion decisions.

STATION3

SMALL BOWEL ANASTOMOSIS

INSTRUCTIONS TO CANDIDATES

You have just resected a segment of small bowel. Perform a single layer, interrupted, end to end anastomosis to restore continuity

ITEM		Not Done or Incorrect	Done Correctly
1.	Bowel oriented mesenteric border to mesenteric border, no twisting	0	1
2.	Stay sutures held with hemostats	0	1
3.	Selects appropriate needle driver (Gen surg, medtip/med or short length)	0	1
4.	Selects appropriate suture (atraumatic, 3.0/4.0, PDS/Dexon/Vicryl/silk)	0	1
5.	Needle loaded 1/2 to 2/3 from tip	0	1
6.	Index finger used to stabilize needle driver	0	1
7.	Needle enters bowel at right angles 80% of bites	0	1
8.	Single attempt at needle passage through bowel 90% of bites.	0	1
9.	Follow through on curve of needle on entrance on 80% of bites	0	1
10.	Follow through on curve of needle on exit on 80% of bites	0	1
11.	Forceps used on seromuscular layer of bowel only majority of time	0	1
12.	Minimal damage with forceps	0	1
13.	Uses forceps to handle needle	0	1
14.	Inverting sutures	0	1
15.	Suture spacing 3 to 5 mm	0	1
16.	Equal bites on each side 80% of bites	0	1
17.	Individual bites each side 90% of bites	0	1
18.	Square knots	0	1
19.	Minimum three throws on knots	0	1
20.	Suture cut to appropriate length (does not interfere with next stitch)	0	1
21.	No mucosal pouting	0	1
22	Apposition of bowel without excessive tension on sutures.	0	1
MAXIMUM TOTAL SCORE			(22)
TOTAL SCORE			

EXAMINER _____

Figure 1. A sample of a task-specific checklist used in the Objective Structured Assessment of Technical Skill (OSATS). Task-specific checklists were used in conjunction with global rating forms at each of the eight stations.

There have been attempts, in the past, to add structure to the assessment of technical skills.[4] Kopta[5] reported high interrater reliability when a checklist approach was used to assess technical competence of orthopaedic trainees. Lossing and Groetzch[6] employed a checklist approach in a multiple station format to assess the efficacy of a course on technical skill given to clinical clerks. Winckle and colleagues[3] from our research group have reported on the reliability and validity of global rating forms and operation specific checklists as tools used in evaluating operative skills by residents in the operating room. Lippert and Farmer[7] have suggested that the acquisition of technical skills is not unidimensional, and have advocated a complete evaluation system to analyze multiple aspects of technical competence.

Although the operating room is intuitively the ideal location for the assessment of technical skills, the routine use of the operating room for systematic and standardized assessment of residents has obvious limitations. First, standardizing any operation has difficulties. Second, standardizing the degree to which a resident is actually performing elements of an operation is almost impossible. Finally, the cost of the "surgical minute" makes the operating room an inappropriate venue for teaching and testing fundamental surgical skills to junior level trainees. Last, and by no means least, is public expectations and awareness of the purpose of the operating room. Indeed, there may be parallels to other professions in which a great degree of technical skill is needed. For example, the airline industry has invested heavily in the development of realistic flight simulators for the purposes of training future pilots and certifying readiness to fly.

Alternatives to the operating room include cadaver models, live animals, and bench model simulations. The use of cadavers has a long tradition in medicine. In particular, the fresh human cadaver has high fidelity to the real world of human operations and simulates tissue dissection well. However, the extreme lack of availability and high costs of procurement and handling rule out this platform as a candidate for routine use for testing. Animal models have been used extensively in medical training. Their use, however, is not universally accepted. The use of animals for purposes of gaining proficiency in surgical skills has been banned in

GLOBAL RATING SCALE OF OPERATIVE PERFORMANCE

Please circle the number corresponding to the candidate's performance in each category, **irrespective of training level.**

	1	2	3	4	5
Respect for Tissue:	Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments		Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissues appropriately with minimal damage
Time and Motion:	Many unnecessary moves		Efficient time/motion but some unnecessary moves		Clear economy of movement and maximum efficiency
Instrument Handling:	Repeatedly makes tentative or awkward moves with instruments by inappropriate use of instruments		Competent use of instruments but occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Knowledge of Instruments:	Frequently asked for wrong instrument or used inappropriate instrument		Knew names of most instruments and used appropriate instrument		Obviously familiar with the instruments and their names
Flow of Operation:	Frequently stopped operating and seemed unsure of next move		Demonstrated some forward planning with reasonable progression of procedure		Obviously planned course of operation with effortless flow from one move to the next
Use of Assistants:	Consistently placed assistants poorly or failed to use assistants		Appropriate use of assistants most of the time		Strategically used assistants to the best advantage at all times
Knowledge of Specific Procedure:	Deficient knowledge. Needed specific instruction at most steps		Knew all important steps of operation		Demonstrated familiarity with all aspects of operation

OVERALL ON THIS TASK, SHOULD THE CANDIDATE: **FAIL** **PASS**

Figure 2. The global rating form used to assess technical skill at each of the eight stations in the Objective Structured Assessment of Technical Skill (OSATS). Global rating forms were used in conjunction with task-specific checklists.

Great Britain since 1876.[8] The extreme efforts of many animal rights groups have served to heighten sensitivity to the continued use of animals for biomedical and educational research. Furthermore, a direct comparison of a live animal platform versus a bench model simulation platform demonstrated the psychometric performance of the two models were equivalent.[2]

One need not simulate a whole operation. In fact, it is better from a testing perspective to break up tasks into their various components, and test each component individually. Our ability, in future, to develop an array of bench model simulations will be limited only by the time and effort invested in the task. With the increasing sophistication of high technology approaches in education, such as virtual reality, a large bank of stations that simulate operative tasks, is well within reach.

For any test to be used with confidence it must possess three qualities. First, it must be feasible and cost effective. Second, it must be reliable. Third, it must be valid. In this experiment, we have demonstrated the feasibility of the OSATS. It is, however, a labor intensive and costly form of examination. The approximate cost per trainee is $200 Canadian. This cost takes into account personnel, materials, and developmental costs but does not take into account costs associated with examiners, as surgeons volunteered their time for this effort.

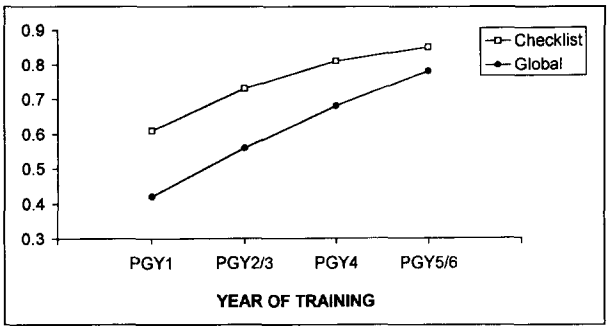

Figure 3. Mean scores for the global rating form and task specific checklists by year of training for 48 general surgery residents who took the Objective Structured Assessment of Technical Skill (OSATS).

Reliability refers to the precision of a test. It answers a basic question: If a test were to be repeated on two successive occasions, and assuming that there was no learning between the taking of the tests, to what extent would the results be identical? Reliability is an index ranging from 0 to 1.0. Tests with reliability estimates in the 0.0 to 0.50 range are so imprecise that it would be difficult to put much weight on the results. Tests with indices from 0.50 to 0.80 are moderately reliable, and tests with reliability indices greater than 0.80 can be used with confidence for high-stakes purposes such as certification. The reliability indices of 0.78 and 0.85 seen in this examination are excellent for a 2-hour examination and give us confidence in interpreting the results.

Validity refers to the concept of whether or not a test measures what it purports to measure. There are many types of validity, such as content, predictive, and construct. An examination cannot be proven valid in any one experiment. Rather, over time and experimentation one accrues evidence for the validity of a test. In this iteration we examined the notion of construct validity. This concept refers to the capacity of a test to measure the trait it was intending to measure, in this case technical skill. One way of gaining evidence for construct validity is to ensure that individuals that are likely to be more competent in the trait perform better on the examination. In this regard, we analyzed the degree to which residents at different levels of training performed at different levels. We were encouraged to see systematic growth in every category of year of training in both approaches to scoring, the global rating and the task-specific checklist.

On the basis of the initial experience, we have adopted the use of an annual test of technical competence for our general surgical training program. After the test, all residents are given a score sheet that details their performance at the various stations, their relation to all test takers, and their relative standing with respect to their peer group by year of training. The results become part of the residents' dossier and are reviewed with the trainee by the program director along with measures of performance in other domains. We feel that there are many potential positive spin-offs to doing regular testing. The residents receive an unambiguous message about the importance of technical skill development. We believe this test will give us the capacity, over time, to identify outliers. As an extension of the identification of outliers, the OSATS may aid in the identification of residents with problems in technical skill at an early time in their training, allowing for the development of timely systematic programs of technical skill enhancement. Finally, the OSATS will enable us to measure the efficacy of programs we are developing for shifting some of the teaching of technical skill out of the operating room and into a laboratory based environment.

REFERENCES

1. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med.* 1990;2:58–76.

2. Martin JA , Regehr G, Reznick RK, et al. An objective structured assessment of technical skill for surgical residents. Presented at the annual meeting of the Society for Surgery of the Alimentary Tract; May 1995; San Diego, Calif.

3. Reznick RK. Teaching and testing technical skills. *Am J Surg.* 1993;165:358–361.

4. Kopta JA. An approach to the evaluation of operative skills. *Surgery.* 1971;70:297–303.

5. Lossing A, Gretsch G. A prospective controlled trial of teaching basic surgical skills with 4th year medical students. *Med Teacher.* 1992;14:49–52.

6. Winckle C, Reznick RK, Cohen R, Taylor BR. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg.* 1994;167:423–427.

7. Lippert FG, Farmer JA. *Psychomotor Skills in Orthopedic Surgery.* Baltimore, Md: Williams & Wilkins; 1984.

8. Cruelty to Animals Act. 15th August 1876, 39 and 40 Vict. Ch77 1-8.