Technical Skills Assessment Toolbox

A Review Using the Unitary Framework of Validity

Iman Ghaderi, MD, Farouq Manji, MD,† Yoon Soo Park, PhD,‡ Dorthea Juul, MD,§ Michael Ott, MD,† Ilene Harris, PhD,‡ and Timothy M. Farrell, MD**

Objective: The purpose of this study was to create a technical skills assessment toolbox for 35 basic and advanced skills/procedures that comprise the American College of Surgeons (ACS)/Association of Program Directors in Surgery (APDS) surgical skills curriculum and to provide a critical appraisal of the included tools, using contemporary framework of validity.

Background: Competency-based training has become the predominant model in surgical education and assessment of performance is an essential component. Assessment methods must produce valid results to accurately determine the level of competency.

Methods: A search was performed, using PubMed and Google Scholar, to identify tools that have been developed for assessment of the targeted technical skills.

Results: A total of 23 assessment tools for the 35 ACS/APDS skills modules were identified. Some tools, such as Operative Performance Rating System (OSATS) and Objective Structured Assessment of Technical Skill (OPRS), have been tested for more than 1 procedure. Therefore, 30 modules had at least 1 assessment tool, with some common surgical procedures being addressed by several tools. Five modules had none. Only 3 studies used Messick's framework to design their validity studies. The remaining studies used an outdated framework on the basis of "types of validity." When analyzed using the contemporary framework, few of these studies demonstrated validity for content, internal structure, and relationship to other variables.

Conclusions: This study provides an assessment toolbox for common surgical skills/procedures. Our review shows that few authors have used the contemporary unitary concept of validity for development of their assessment tools. As we progress toward competency-based training, future studies should provide evidence for various sources of validity using the contemporary framework.

Keywords: assessment, competency, surgical education, technical skills, toolbox, unitary framework of validity

(*Ann Surg* 2015;261:251–262)

Over the past decade, duty-hour restrictions have driven reevaluation of Halstead's traditional apprenticeship model for surgical training. In the Halstedian approach, trainees achieved competency by performing a large numbers of surgical cases.[1] This training model required a significant dedication of time and sacrifices in trainees' personal lives as they spent long hours in the hospital to perform an adequate number of procedures. Gradually, concern was raised regarding the intertwined issues of resident fatigue, medical errors, and patient safety.[2] In parallel, there has been a generational shift in trainees' attitudes about life-work balance.[3,4] In 2003, the Accreditation Council for Graduate Medical Education (ACGME) mandated a restriction in trainee duty hours in all specialties. Philibert and colleagues[5] commented, "The aim of these standards was to promote high-quality learning and safe care in teaching institutions." In addition, in 2008, the Institute of Medicine released its report on resident duty hours[6] and recommended additional restriction to duty hours and other changes in training programs to enhance the experience for residents and improve patient safety.

With implementation of duty-hour regulations, concerns were raised that trainees might not have adequate time to develop competencies in the required surgical skills. It has been shown that the 80-hour workweek would result in 6 months to a year reduction of in-hospital experience in a 5-year residency program. The overall in-hospital experience, including management of urgent and emergent conditions, has been considerably impacted with an estimated reduction of 33% to 50%. This reduction in overall in-hospital experience also includes a decrease in the opportunities for assisting in surgeries, which is an important component of surgical training.[7] Currently, one of the challenges facing surgical residency programs is providing adequate cognitive and technical training to achieve competency levels during the course of residency.

In 1999, the ACGME introduced a requirement for instruction and assessment in 6 domains of clinical competency. A decade later, the ACGME began a process of restructuring its accreditation system on the basis of assessment of these competencies. This system is called the "Next Accreditation System". A key element of the Next Accreditation System is assessment of competencies referred to as educational milestones. The milestones are "developmentally based, specialty specific achievements that residents are expected to demonstrate at established intervals as they progress through training."[8] The aim was to "create a logical trajectory of professional development in essential elements of competency and meet criteria for effective assessment."[9] Accordingly, surgical educators and leaders have begun to consider approaches to increase the effectiveness of surgical education. Modification of current curricula can be a stepping stone.

In 2005, the American College of Surgeons (ACS) and the Association of Program Directors in Surgery (APDS) formed the Surgical Skills Curriculum Task Force to design a national skills curriculum to enhance training of surgical residents, using a simulated environment to better prepare trainees for performance in the operating room.[10] The ACS/APDS curriculum was designed on the basis of ACGME core competencies and consists of 3 phases (Table 1). Phase 1 includes 20 basic surgical skills modules; phase 2 includes 15 advanced surgical skills modules; and phase 3 includes 10 team-based skill modules.[11] This curriculum has criterion-based goals and requires trainees to complete modules to a proficiency level before performing any procedure in the operating room.

Program evaluation is an important step in curriculum implementation and involves evaluating the effectiveness of a given

From the *Department of Surgery, The University of North Carolina at Chapel Hill, Chapel Hill, NC; †Department of Surgery, Western University, London, Canada; ‡Department of Medical Education, University of Illinois at Chicago, Chicago, IL; and §American Board of Psychiatry and Neurology, Inc., Deerfield, IL.

This study was presented at 2013 Surgical Education week, Orlando, Florida.

Disclosure: The authors received no funding for this study. The authors declare no conflicts of interest.

Reprints: Iman Ghaderi, MD, or Timothy M. Farrell, MD, Department of Surgery, The University of North Carolina at Chapel Hill, 4035 Burnett-Womack Bldg, CB #7081, Chapel Hill, NC 27599. E-mail: iman.ghaderi@gmail.com or timothy_farrell@med.unc.edu.

ISSN: 0003-4932/14/26102-0251

DOI: 10.1097/SLA.0000000000000520

TABLE 1. Phases 1/2/3: Basic/Core, and Advanced Skills and Tasks, and Team-based Training Modules

No.	Phase 1 Modules
1	Asepsis and instrument handling
2	Knot tying
3	Suturing
4	Tissue handling, dissection, wound closure
5	Advanced tissue handling, flaps and skin grafts
6	Catheterization, urethral and suprapubic
7	Airway management
8	Chest tube and thoracentesis
9	Central venous access, arterial lines
10	Surgical biopsy
11	Arterial anastomosis
12	Laparotomy, opening and closure
13	Principles of bone fixation and casting
14	Inguinal anatomy
15	Upper endoscopy
16	Lower endoscopy
17	Basic laparoscopic skills
18	Advanced laparoscopic skills
19	Handsewn anastomosis
20	Stapled anastomosis
	Phase 2 Modules
1	Laparoscopic ventral hernia repair
2	Laparoscopic colon resection
3	Laparoscopic/open bile duct exploration
4	Abdominal wall stomas
5	Laparoscopic appendectomy
6	Laparoscopic Nissen fundoplication
7	Sentinel node biopsy and axillary lymph node dissection
8	Open inguinal/femoral hernia repair
9	Laparoscopic inguinal hernia
10	Laparoscopic/open splenectomy
11	Laparoscopic/open cholecystectomy
12	Thyroidectomy
13	Parathyroidectomy
14	Gastrectomy
15	Distal/total pancreatectomy
	Phase 3: Team-Based Skills
1	Laparoscopic crisis
2	Laparoscopic troubleshooting
3	Latex allergy anaphylaxis
4	Patient handoff
5	Postoperative hypotension
6	Postoperative MI (cardiogenic shock)
7	Postoperative pulmonary embolus
8	Preoperative briefing
9	Retained sponge on postoperative chest radiography
10	Trauma team training

curriculum in increasing participants' knowledge and skills in targeted areas.[12] Therefore, an assessment phase was included by the ACS/APDS to ensure that residents achieve proficiency by the end of training in the new curriculum. Thus, assessment of trainees is an essential component of program evaluation.

Recently, the American Board of Surgery (ABS) mandated a new requirement for assessment of operative and clinical performance for candidates who will apply for the general surgery certification examination. Applicants are required to obtain 2 operative performance assessments and 2 clinical performance assessments conducted by their program director or another faculty member during residency training. According to the ABS, this requirement will increase to 6 operative performance assessments and 6 clinical performance assessments for graduates who will complete their training in the next few years.[13]

This paradigm shift toward competency-based training has made assessment an important research subject in the field of surgical education. Traditionally, technical skills of trainees have been assessed using in-training evaluation reports that are completed by faculty. Although these assessments, which reflect an expert's opinion of trainee performance, are valuable, they may not produce valid results.[14] They are generally subjective and usually fail to provide specific suggestions to improve deficiencies.[15]

In previous literature reviews,[16,17] authors have organized their reviews focused on assessment instruments rather than procedures. Instrument-based reviews are not helpful to an end user who is simply looking for an appropriate assessment tool for a specific procedure. There is a need for an easily navigated assessment toolbox for users such as program directors who want to assess their residents' competency in specific surgical procedures.

The purpose of this article is to provide a technical skills assessment toolbox, with critical appraisal of the selected tools using Messick's unitary framework of validity. Thirty-five basic and advanced skills/procedures, which mirror the ACS/APDS surgical skills curriculum, were chosen as the competencies for assessment instruments in this toolbox. This review also provides insight into the areas where gaps exist, and further research and development is needed for competency-based assessments.

VALIDITY

In the current state-of-the-art conceptual framework, validity is defined as appropriate interpretation of test results, and a validation study is a process of collecting evidence to support the interpretations of assessment results.[18] In the traditional framework, validity consists of 3 separate types: content, criterion (including concurrent and predictive validity), and construct.[19,20] Messick,[21] however, argues that "the traditional concept of validity is fragmented and incomplete, failing to take into account evidence of the value implications of score meaning as a basis for action and of the social consequences of score use." Instead, he offers a unitary concept of validity that

> interrelates these issues as fundamental aspects of a more comprehensive theory of construct validity addressing both score meaning and social values in test interpretation and use and integrating content, criterion, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and relevant relationships.[21]

Messick[21(p741)] emphasizes that

> the construct validation [of] the test score is not equated with the construct it attempts to tap, nor is it considered to define the construct. [. . .] Rather, the measure is viewed as just one of an extensible set of indicators of the construct.

Therefore, validity applies only to the scores or interpretation, in a specific context, and the commonly used term "valid instrument" is inaccurate.[22–24] In addition, "because validity is a property of inferences, not instruments, validity must be established for each intended interpretation."[22] This principle is based on Kane's argument-based approach to validation, indicating that the validity of the results should be established for each assessment.[25,26] For example, if an instrument can produce valid scores assessing performance in a simulator, it cannot be used for assessment of the same procedural skill in the operating room without a complementary validation study for the proposed procedure. Thus, "a clear definition of the intended construct is the first step in any validity evaluation."[25]

In 1999, the unitary definition of validity was endorsed by the American Educational Research Association, the American

Psychological Association, and the National Council on Measurement in Education (Standards 1.1, 1.2, 1.3, and 1.4).[18] Since then, this definition has been incorporated in the Standards for Educational and Psychological Testing. In the contemporary framework, *construct* validity is the only form of validity.[20] Researchers investigate the "evidence for the validity of results and the use of those results" from multiple sources.[20] In the new approach, validity is a construct with various facets, and the validation process requires identification of the relevant "sources of validity" for these facets. Therefore, the phrase "types of validity" has been abandoned and replaced with "sources of validity."[18] These sources are content, response process, internal structure, relationships to other variables, and consequences of testing.[18] A brief description of each source of validity, with relevant examples, is provided in Table 2.

Over the past decade, many tools for assessment of performance in a wide variety of procedures have been designed. Although the unitary framework of validity was introduced more than a decade ago, the majority of studies have applied the traditional framework for validity to design their assessment tools. Similarly, most published systematic reviews[15,27–29] have used the traditional framework of validity to examine the quality of assessment tools, with the authors reporting on the different "types of validity" (face, content, construct, concurrent, predictive validity).

This use of outdated concepts of validity demonstrates a lack of familiarity with the contemporary framework accepted as superior by researchers and authors in surgical education. In 2010, Korndorffer et al[30] called for use of the contemporary definition of validity in the surgical literature. They reviewed validation studies in laparoscopic simulator education and demonstrated the limited use of the contemporary framework for establishing validity. They suggested that surgical educators should use this framework to examine assessment methods to avoid inappropriate assessment of performance.[30]

In 2 recent systematic reviews, Van Hove et al[16] and Ahmed et al[17] identified and reviewed the instruments that have been developed for objective assessment of procedural skills. Van Hove et al used the review by Gallagher et al[31] that was published in 2003, and Ahmed et al used the definitions proposed by Van der Vleuten[32] for validity. These authors used "types of validity," that is, the traditional framework, in their articles. Van Hove et al also evaluated the studies by applying the evidence-based medicine levels of evidence. Although the authors of the aforementioned reviews used the traditional framework for conceptualizing validity, they offered

TABLE 2. Validity: Sources of Evidence*

Evidence Source	Definition	Examples
Content	The "relationship between a test's content and the construct it is intended to measure."	Test blueprint Representativeness of items to the domain Logical/empirical relationship of content tested to achievement domain Development strategies to ensure appropriate content representation Item writer qualifications Analyses by experts for adequacy of items representing the content domain
Process response	Analyses of responses (actions, strategies, thought processes) of individual respondents or observers. Differences in response processes may reveal sources of variance irrelevant to the construct being measured. It includes instrument security, scoring, and reporting of results.	Trainee format familiarity Understandable/accurate descriptions/interpretations of scores for trainees Rater training Quality control of scoring Validation of preliminary scores (pilot study) Accuracy in combining different format scores Quality control/accuracy of final scores/marks/grades Subscore/subscale analyses Accuracy of applying pass–fail decision rules to scores
Internal structure	Degree to which individual items within an instrument fit the underlying constructs. It is often reported by measures of internal consistency reliability and factor analysis.	Item analysis data [item difficulty/discrimination, item/test characteristic curves (ICCs/TCCs), interitem correlations, item-total correlations] Score scale reliability Generalizability Item factor analysis Psychometric model
Relations to other variables	Relationship between scores and other variables relevant to the construct being measured. Relationships may be positive (convergent/predictive) or negative (divergent/discriminant) depending on the constructs being measured.	Correlation with other variables or scores on other performance assessments (correlation between postgraduate level and scores) Test-criterion correlations Generalizability of evidence
Consequences	Assessments are intended to have some desired effect or may have unintended effects.	Impact of test scores/results on trainees Consequences for learners/future learning Positive consequences outweigh unintended negative consequences? Reasonableness of method of establishing pass–fail (cut) score Pass–fail consequences (P/F decision reliability–classification accuracy) Instructional/learner consequences Method of determining pass–fail score; differential pass–fail rates among examinees expected to perform similarly

*From Downing and Yudkowsky[49] and Beckman et al.[34]

valuable information about the current status of research on assessment of technical performance and the validation process in the field of surgical education.

METHODS

A search was performed using PubMed and Google Scholar to identify the instruments that have been developed for assessment of targeted technical skills. The reference lists of previous systematic reviews were used as the benchmark for the search results. The selected studies were reviewed, using the unitary framework of validity to interpret their results. The evidence authors provided for validity of their interpretations concerning assessed constructs was reviewed. These studies were examined to determine whether their results could be interpreted according to the new conceptualization of validity. Studies were evaluated on the basis of whether they evaluated different sources of validity using the current unitary concept, that is, "construct validity" or "types of validity" and whether the authors reported a "valid instrument" as distinguished from "valid scores/interpretations."[33]

As the majority of studies used the traditional framework, we extracted data that could be considered sources of validity. To quantify the degree each source of validity was reported, we adopted the rating system of Beckman et al,[34] with some modifications. Beckman and colleagues originally rated studies as follows: "N" for studies with no discussion of the source of validity evidence and/or no data presented, "0" for studies that discussed the source of validity evidence but did not provide any data or the data failed to support the validity of instrument scores, and "1" or "2" for studies in which the data "weakly" or "strongly" supported the validity of score interpretations, accordingly. Beckman and colleagues did not award a score of "0" to any study in the *content* category. They stated that "any discussion of content evidence would constitute data, and it seems unlikely that data would not at least weakly support the content of any published instrument."[34]

An issue with Beckman's rating scale is that it is essentially dichotomous, with studies either weakly or strongly supporting the validity of score interpretations. The advantage of this simplified scale is its ease of use and high agreement among reviewers. On the contrary, it does not represent the wide variation that studies have provided for the validity of their interpretations. Therefore, we broadened the above rating scale and added an extra level. Our scoring system was as follows: 0 = no discussion or data presented as a source of validity evidence; 1 = data that weakly support the validity of score interpretations; 2 = some data (intermediate level) that support the validity of score interpretations, but with gaps; and 3 = multiple sets of data that strongly and completely support the validity of score interpretations (Table 3).

The full texts of relevant articles were retrieved and authors I.G. and F.M. independently reviewed the selected studies and assigned scores using the above rating scale. The interrater reliability between their scores was calculated using intraclass correlations. Disagreements were resolved by discussion to reach consensus on the final ratings (Table 5).

The ACS/APDS curriculum skills modules (all 35 procedural modules in phases 1 and 2) were used as a template to create the toolbox. Although their curriculum was originally proposed to prepare residents before entering the operating room, assessment of skills and determination of competency within a training program are not restricted to a simulated environment. Thus, we included all the relevant tools regardless of whether they have been developed and tested in the laboratory or the clinical workplace (operating room or endoscopy suite). Using a similar template in both the simulated environment and the operating room creates a universal standard that enables programs to monitor the performance of residents longitudinally. The only caveat is that these assessment tools should produce valid results in the operating room before their formal implementation.

We did not include motion analysis systems, such as ICSAD,[35] ADEPT,[36] ProMIS,[37] HUESAD,[38] and TrEndo,[39] in our toolbox. These instruments have been used mostly in simulation laboratories and occasionally in operating rooms in the institutes where they were developed. They require extensive infrastructure, and implementation of these systems is costly and is, therefore, generally not feasible.

With regard to the simulators, we included only the FLS[40] (Fundamentals of Laparoscopic Surgery) curriculum in our toolbox. Although the FLS laparoscopic trainer box is designed for training and assessment of laparoscopic skills in a simulated setting, it has been shown to correlate positively with performance in the operating room.[41–44] For these reasons, the FLS curriculum is now part of the basic and advanced laparoscopic skills modules included in the ACS/APDS National Skills Curriculum and is one of the requirements for certification by the ABS.

RESULTS

Twenty-three assessment instruments for 30 modules (out of 35) were identified. We identified no instruments for assessment of 5 procedures/skills: Advanced tissue handling; flaps and skin grafts; catheterization; urethral and suprapubic, inguinal anatomy; laparoscopic/open splenectomy; and distal/total pancreatectomy. Table 4 describes the characteristics of these studies and the reported instruments. It also includes the setting (operating room [OR] or endoscopy suite vs simulated environment), the framework of validity (valid tool vs valid results/interpretations), and the presence of evidence for sources of validity. Only 3 studies (OPRS,[45] Mayo Colonoscopy Skills Assessment Tool [MCSAT],[46] Ottawa Surgical Competency Operating Room Evaluation [O-SCORE][47]) used Messick's framework to design and interpret their results. The rest used the traditional framework of validity. When these results were analyzed in the unitary framework, a few sources of validity, including content, internal structure, and relationship to other variables were identified. A majority of these studies failed to provide evidence for the consequences of assessments. Some tools, such as OSATS,[48] OPRS,[45] and O-SCORE,[47] have been tested for more than 1 procedure (modules 4, 8, and 13), and more than 1 assessment instrument has been developed for some common surgical procedures such as laparoscopic cholecystectomy, laparoscopic colectomy, and hernia repair. The 2 main measures that have been reported are reliability (interrater reliability and/or internal consistency) and whether these tools generated scores that could differentiate trainees with different levels of skills. The latter has been assumed as adequate evidence for construct validity by many authors. In the contemporary framework, these 2 aspects can be categorized under internal structure and relations to other variables, respectively. The overall agreement between 2 raters was high (83.64%). Interrater reliability was good for all categories of validity (weighted κ range: 0.68–0.88) (Table 5).

DISCUSSION

Our results show that there has been a considerable lag in use of the contemporary framework for conceptualizing validity, and that framework has not been adopted by surgical educators. It seems that over the past decade, assessment in surgical literature has evolved in isolation, without benefit from advances in the science of assessment in parallel fields. In this section, the results of the aforementioned studies are analyzed on the basis of the 5 sources of validity evidence specified in the unitary framework.

TABLE 3. Criteria for Rating Validity Evidence

Evidence Category	Rating*	Rating Criteria
Content	0	No discussion or data regarding the instrument content
	1	Only discussion or limited amount of data (simply listing items without justification)
	2	Listing assessment themes with some references and justifications, limited description of the process for creating the instrument Alternatively, reference to a prior study on an assessment instrument that meets these criteria
	3	Well-defined process for developing instrument content, including both an explicit theoretical/conceptual basis for instrument items and systematic item review by experts
Response process	0	No discussion or data regarding the response process
	1	Minimal discussion and limited data presented. Use of an instrument without reporting the results. Discussing the impact of response rate on assessment scores or speculating on the thought processes of learners
	2	Some data regarding thought processes and analysis of responses. Some data about implication of systems that reduced response error
	3	Multiple sources of supportive data, including critical examination of thought processes, analysis of responses for evidence of halo error or rater leniency, or data demonstrating low response error
Internal structure	0	No discussion or data
	1	Minimal data with regard to internal structure, some reliability with a single measure
	2	Factor analysis incompletely confirming anticipated data structure or a few measures of reliability reported
	3	Factor analysis confirming anticipated data structure or multiple measures of reliability. Item analysis data, item/test characteristic curves (ICCs/TCCs), interitem correlations, item-total correlations), generalizability analysis
Relation to other variables	0	No discussion or data
	1	Correlation of assessment scores to outcomes with minimal theoretical importance, a single measure of validity (relationship between level of training and scores)
	2	Correlation of assessment scores to outcomes with some theoretical importance
	3	Correlation (convergence) or no correlation (divergence) between assessment scores and theoretically predicted outcomes or measures of the same construct. Such evidence will usually be integral to the study design and anticipated a priori, generalizability evidence
Consequences	0	No discussion or data
	1	Limited data about the consequences of the assessment. Merely discussion about the consequences of assessment (eg, data regarding usefulness of assessment based on postassessment survey)
	2	Description of consequences of assessment that could conceivably impact the validity of score interpretations (although these impacts are not explicitly identified by the authors)
	3	Description of consequences of assessment that clearly impact on the validity of score interpretations, as supported by data and convincingly argued by the authors. Such evidence will usually be integral to the study design and anticipated a priori

*Scoring system: 0 = studies that no discussion or data presented as a source of validity evidence; 1 = the study provided data that weakly support the validity of score interpretations; 2 = the study provided some data (intermediate level) that support the validity of score interpretations but it was incomplete; and 3 = the study provided multiple sets of data that strongly supported the validity of score interpretations.

Adopted and modified from Beckman et al.[34]

Content

Content evidence refers to the "relationship between test content and the construct of interest."[49] To provide evidence for this aspect of validity, an educator or practitioner who has expertise in the related content domain would create a blueprint that is representative of a targeted construct (eg, surgical skill/ procedure). There should be a "logical and empirical" relationship between the content of the test and the content domain of the construct.[49]

With respect to this source of validity, the authors of the majority of these studies mention, to some extent, the process of developing their assessment instruments. The expert consensus approach was the most common method. In this approach, local experts created a blueprint, using their own expertise and input from textbooks, videos of procedures, and pertinent literature. There was, however, a large variation with respect to use of this methodology in the development of the assessment tools and the amount of information authors reported. Although authors of some studies provided very little information about this process,[50] others, such as Palter et al,[51] carried out an extensive process, using the Delphi model to incorporate the opinions of a larger pool of experts in different countries to ensure the comprehensiveness of their blueprint for laparoscopic colectomy. Some authors included independent content experts to review their blueprint or used a multicenter design to improve the representativeness of their tool, including experts across several institutions.[52,53]

Response Process

Response process is defined as "evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible."[20] It entails analysis of responses and accuracy of scoring and reporting of results. Because the differences in response processes may result in variance that is irrelevant to the construct being measured, one must examine "the reasoning and thought processes of learners or systems in order to reduce the likelihood of response error."[34] Therefore, for example, instrument items and anchors describing points on the rating scale

TABLE 4. Technical Skills Assessment Toolbox

				Concept of Validity		Source of Validity				
	Procedure	Tool	Setting	The Tool	The Results	Content	Response Process	Internal Structure	Relations to Other Variables	Consequences
1	Asepsis and instrument handling	Chipman and Schmitz[60] (OSATS)	Laboratory (simulation)	+		2	1	2	1	1
2	Knot tying	Chipman and Schmitz [60] (OSATS)	Laboratory (simulation)	+		2	1	2	1	1
		Scott et al[61]	Laboratory (simulation)	+		1	2	1	1	1
		Swift and Carter (OSATS)[62]	Laboratory (simulation)	+		2	2	2	2	1
3	Suturing	Chipman and Schmitz[60] (OSATS)	Laboratory (simulation)	+		2	1	2	1	1
4	Tissue handling, dissection, wound closure	Shippey et al[63] (modified OSATS)	Laboratory (simulation)	+		2	1	2	1	0
		Chipman and Schmitz[60] (skin closure)	Laboratory (simulation)	+		2	1	2	1	1
5	Advanced tissue handling, flaps, and skin grafts	Ø								
6	Catheterization, urethral and suprapubic	Ø								
7	Airway management	O'Conner and McGraw[64]	Laboratory (simulation)	+		2	1	1	0	0
8	Chest tube and thoracentesis	Anastakis et al[65] (OSATS)	Laboratory (cadaveric)	+		2	1	2	1	1
9	Central venous access, arterial lines	OPRS[45]	OR		+	2	1	2	2	0
		GRITS[66]	OR	+		2	1	2	1	0
10	Surgical biopsy	OPRS[45] (excisional biopsy)	OR		+	2	2	3	2	3
11	Arterial anastomosis	Wilasrusmee et al[67,68] (modified OSATS)	Laboratory/OR	+		1	1	1	2	1
12	Laparotomy, opening and closure	O-SCORE[47]	OR		+	2	2	3	2	2
		Anastakis et al[65] (OSATS)	Laboratory (cadaveric)	+		2	1	2	1	1
13	Principles of bone fixation and casting	Leong et al[69] (MOSATS)	Laboratory (animal)	+		1	2	2	2	0
14	Inguinal anatomy	Ø								
15	Upper endoscopy	Cass et al[70]	Endoscopy suite	+		1	2	0	1	1
		Chack et al[71]	Endoscopy suite	+		2	1	0	2	1
		GAGES[52]	Endoscopy Suite	+		2	2	2	2	2

(*continued*)

TABLE 4. (*Continued*)

				Concept of Validity		Source of Validity				
	Procedure	Tool	Setting	The Tool	The Results	Content	Response Process	Internal Structure	Relations to Other Variables	Consequences
16	Lower endoscopy	Cass et al[70]	Endoscopy suite	+		1	2	0	1	1
		Chack et al[71]	Endoscopy suite	+		2	1	0	2	0
		GAGES[52]	Endoscopy Suite	+		2	2	2	2	2
		MCSAT[46]	Endoscopy suite		+	2	3	3	2	2
17	Basic laparoscopic skills (FLS, 1–3 tasks)	FLS[40]	Laboratory (simulation)	+		3	3	3	3	3
18	Advanced laparoscopic skills (FLS 4, 5 tasks, continuous suturing)	FLS[40]	Laboratory (simulation)	+		3	3	3	3	3
		Van Sickle et al[50]	OR/Animal	+		2	2	2	1	1
19	Handsewn anastomosis	Datta et al[72] (modified OSATS)	Laboratory (simulation)	+		2	2	2	2	0
		Shah et al[73] (OSATS)	Laboratory (simulation)	+		2	2	2	2	1
		Bann et al[74] (OSATS)	Laboratory (simulation)	+		2	2	1	1	0
		Munz et al[75] (modified OSATS)	Laboratory (simulation)	+		2	2	2	1	0
20	Stapled anastomosis	Vick et al[76]	Laboratory (simulation)	+		1	1	0	1	0
21	Laparoscopic ventral hernia repair	GOALS-IH[44]	OR/laboratory	+		2	2	2	2	0
		OPRS[45]	OR		+	2	1	2	2	0
22	Laparoscopic colon resection	Sidhu et al[77] (modified OSATS)	Animal laboratory	+		2	1	2	1	1
		Palter and Grantcharov[78]	OR	+		3	2	2	2	0
		Sarker et al[79]	OR (videotaped)	+		2	1	2	1	1
		Miskovi et al[80]		+		2	1	3	1	2
23	Laparoscopic/open bile duct exploration	Santos et al[81]	Laboratory	+		2	2	2	1	0
24	Abdominal wall stomas	Szalay et al[82]	Laboratory (simulation)			1	1	2	1	0
25	Laparoscopic appendectomy	GOALS[83]	OR	+		3	3	3	3	1
		O-SCORE[47]	OR		+	2	2	3	2	2

(*continued*)

TABLE 4. (*Continued*)

				Concept of Validity		Source of Validity				
	Procedure	Tool	Setting	The Tool	The Results	Content	Response Process	Internal Structure	Relations to Other Variables	Consequences
26	Laparoscopic Nissen fundoplication	Modified OSATS[84]	Laboratory (animal)	+		2	1	1	1	0
27	Sentinel node biopsy and axillary lymph node dissection	OPRS[45]	OR		+	2	1	2	2	0
		O-SCORE[47]	OR		+	2	2	3	2	2
28	Open inguinal/ femoral hernia repair	OPRS[45]	OR		+	2	1	2	2	0
		GRITS[66]	OR		+	2	1	2	1	0
29	Laparoscopic inguinal hernia	GOALS-GH[55]	Laboratory/ OR	+		2	2	2	2	0
		OPRS[45]	OR		+	2	1	2	2	0
		GRITS[66]	OR	+		2	1	2	1	0
30	Laparoscopic/open splenectomy	Ø								
31	Laparoscopic/open cholecystectomy	GOALS[85]	OR	+		3	3	3	3	1
		O-SCORE[47]	OR		+	2	2	3	2	2
		GRITS[66]	OR	+		2	1	2	1	0
		OPRS[45]	OR		+	2	1	2	2	0
32	Thyroidectomy	GRITS[66]	OR	+		2	1	2	1	0
		OPRS[45]	OR		+	2	1	2	2	0
33	Parathyroidectomy	OPRS[45]	OR		+	2	1	2	2	0
34	Gastrectomy	SATOS[86]	Animal/laboratory	+		1	1	0	1	1
35	Distal/total pancreatectomy	Ø								

OR indicates operating room.

TABLE 5. Interrater Reliability by Validity Evidence*

Validity Evidence	% Agreement	ICC
Content	81.82	0.68 (0.17)
Response process	81.82	0.77 (0.17)
Internal structure	78.79	0.79 (0.17)
Relationship to other variables	90.91	0.88 (0.17)
Consequences	84.85	0.85 (0.17)
Overall	83.64	0.85 (0.08)

*Values within parentheses represent standard errors.
ICC indicates intraclass correlations.

should be explicit and clear. Response process also includes accuracy of data collection and the process of data entry into a database.

All of the tools included in this toolbox with the exception of FLS are observational and in the form of global rating scales, checklists, or error-based systems that require an observer to complete the assessment. In the FLS system, the time and preset errors are used as rating measures. The observer could be a practicing surgeon, educator, a faculty member, or a member of the research group. To increase the consistency of observations, observers need specific training about the tool to use it uniformly when assessing performance of trainees. For example, one of the initial steps in any laparoscopic procedure is port placement. The attending surgeon may point out where ports should be placed instead of allowing the trainee to decide independently about appropriate port placement. Giving the trainee the opportunity to decide about port placement does not mean that the attending will allow inappropriate placement of ports, which would compromise patient safety, but rather that he would allow the trainee to demonstrate his or her knowledge of the steps of the operation and the associated relevant skills. The MCSAT tool[46] provides a similar example during assessment of performance in colonoscopy. The author comments, "[During colonoscopy], instructing staff [should] not point out pathology until fellows have had a chance to independently identify or correctly interpret pathologic findings, as is typically done."

The Likert-format rating scale is the main format that has been used in global rating scales. One of the issues with this format is the tendency of assessors to not use the entire scale (scale shrinkage). Although the anchors are written to aid the assessment of technical competencies in a particular task/procedure (criterion-based assessment), the general tendency is to rate the performance of trainees according to their postgraduate year level (normative-based assessment). Therefore, training raters is very important to improve the accuracy of rating, and thereby increase the interrater agreement (discussed under the section "Internal Structure").

Once the validation process commences, collected assessments should be reviewed after a few scoring sessions to verify the accuracy of ratings. This can be done in the form of a pilot study in which issues related to implementation of the assessment form are addressed. An appropriate adjustment to the tool or training of observers can take place accordingly.

Internal Structure

This source of validity evidence "relates to the statistical or psychometric characteristics of" the assessment tool. It is usually referred to as "reliability" and includes reproducibility and generalizability of results.[49] If test scores are not reproducible (ie, not reliable), it is "nearly impossible to interpret the meaning of those scores." In other words, lack of reliability equals lack of validity, and, therefore, reliability is categorized under validity in the unitary framework of validity.

The commonly used measures of this aspect of validity derive from item analysis data (interitem and item-total correlation), and internal consistency and interrater reliability data. With respect to interitem correlation, high correlations between items demonstrate that they measure the same construct. On the contrary, low correlations between items indicate that they may be measuring different constructs review. The majority of the studies assessed interrater (observer) and/or intraclass correlations using Pearson correlation coefficients. Few studies reported internal consistency using Cronbach α.

Only 2 studies used generalizability theory to study reliability. The Southern Illinois University group used this analysis to estimate dependability indices across OPRS ratings.[45] These authors provided an estimate of the number of ratings required to achieve the recommended level of reliability (2.3 OPRs per month). The authors of O-SCORE[47] performed an analysis of variance on the ratings and demonstrated that it would take at least 5 O-SCORE observations per trainee to produce a g-coefficient of 0.80.

Relationship to Other Variables

This aspect of validity is conceptualized as the correlation between the instrument assessment scores and the scores of appropriate criterion measures.[20] Usually, the newer measure is validated against a well-known existing measure, against which both convergent and divergent correlations are possible.[20] The majority of these studies are hypothesis driven. One of the most common correlations, referred to as "known group construct validity,"[54] is that between performance scores and the level of experience or training. Most validation studies in the surgical literature examined assessment tools in 2 groups of subjects expected to have a large gap in skill levels, usually novices and experts. Although this type of correlation represents a small facet of validity, it seems that many authors considered this to be the main source of construct validity and, therefore, made no further attempts to investigate other sources of validity for their tools.

Other studies showed the correlation between the scores of a test and scores generated by other instruments. For example, the McGill group showed that there was a positive correlation between the results of their procedure-specific GOALS (GOALS-IH,[44] GOALS-GH[55]) and the generic GOALS. Datta et al[35] showed a significant correlation between the scores generated using OSATS and that of motion analysis assessments (ICSAD) during standardized laboratory-based tasks. Similarly, Moorthy et al[56] demonstrated a strong correlation between ICSAD scores and total checklist scores. This checklist was developed for assessment of intracorporeal suturing.

Consequences

Another important measure of validity focuses on the consequences of assessment for learners. This category is the least studied aspect of validity.[20] Consequences can be positive or negative and intended or unintended.[49] Messick[21] recommends including "evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation or with unfairness in test use." The significance of consequences depends on whether the tool is designed for formative assessment (for feedback) or summative assessment (pass/fail) and whether it is used for low versus high stakes assessments (eg, some end of rotation evaluations vs yearly promotion or licensing purposes).

Using consequences as a source of validity evidence has been a subject of controversy. Opponents argue that the consequences of an assessment are beyond the scope of a validity study and defer to policy makers the decisions about the impact and the appropriateness of its use.[57] Mehrens comments,[58]

> The consequences of a particular use do not necessarily inform us regarding either the meaning of a construct or the adequacy of a particular assessment process in measuring that construct. Indeed, the meaning of the construct and evidence that the test measures that construct may be well established prior to some specific use.

He adds that the consequences are often political value judgments that may not provide any information about "the accuracy of the inferences about whether the assessment is a good measure of a construct."[58]

In contrast, advocates of including consequences in instrument validation argue that consequences reflect "the soundness of test-based decisions."[57] They believe that the consequences of an assessment can reflect flaws in the conceptualization of the assessment tool and interpretation of the scores/results. For example, the unintended consequences can be due to the presence of threats to validity, for example, "construct underrepresentation or inclusion of sources of construct-irrelevant variance."[57] Other authors, such as Shepard,[59] comment that consequences should be carefully studied but that they categorically are not part of validity.

We believe that the consequences of assessment in competency-based medical and surgical training are crucial, and regardless of whether we consider them a source of validity evidence or not, the consequences of each assessment should be carefully examined, and the evidence for its appropriateness of use must be demonstrated. Among the assessment tools included in this toolbox, only 5 studies (FLS,[40] OPRS,[45] O-SCORE,[47] MCSAT,[46] and GAGES[52]) reported data regarding the consequences of their use. The authors of FLS and OPRS further established the evidence for the consequences of their assessments through several follow-up studies. Currently, the FLS assessment system is used by the ABS as part of its certification process.

In summary, the 2 main validity measures that have been reported are reliability (interrater reliability and/or internal consistency) and discrimination (whether these tools generated scores that could differentiate trainees with different levels of skills). The latter has been assumed to be adequate evidence for *construct validity* by many authors. In the contemporary framework, these 2 aspects can be categorized under internal structure and relationship to other variables, respectively. There are very few studies that have gone beyond these 2 measures and investigated the evidence for other sources of validity.

CONCLUSIONS

This study provides an assessment toolbox for common surgical skills/procedures, with appraisals of the validity evidence for instruments in the toolbox. Our review shows that few authors have utilized the contemporary unitary concept of validity for development and appraisal of their assessment tools. With the current level of validity evidence, we recommend that most assessment tools should be used only for instructional purposes and/or formative assessment. Before any further application of these tools, for summative and high stakes assessments, such as medical certification and recertification, extensive validation processes must first take place. As part of this process, researchers should determine the short- and long-term impact of the results generated by these tools on trainees. These outcomes can be used to establish evidence-based pass–fail scores. In this context, application of generalizability theory can assist researchers to determine how many assessments are needed to obtain an accurate measure of ability.[49] Using this rigorous method, researchers will be able to establish different pass–fail cut points for various levels of performance in different contexts. Ideally, such cut points can be used as a complementary tool to decide whether and when trainees are ready for the OR, after practicing in a simulated curriculum, or whether they should pass a rotation or get promoted at the end of a postgraduate year. Eventually, such thresholds can be used as part of a multisource competency-based assessment for graduation of surgeons from residency programs. For example, FLS certification is one of the ABS requirements for board eligibility.

As we move toward the competency-based training and assessment model, future studies of the assessment instruments should provide evidence for all sources of validity (especially consequences), address the lack of data for generalizability of current assessments, and determine the appropriateness of these methods and instruments for formative and summative assessment. These studies should focus on filling gaps by providing further validity evidence for existing assessment tools. The results from a decade of research in developing assessment tools can provide a platform and foundation for future research. As demonstrated in this study, there are flaws in the design and conceptualization of some of these tools. Future studies should improve existing tools and take advantage of work already done, instead of "reinventing the wheel" by creating new tools where tools already exist.

Although the feasibility of implementing these assessment tools in any training program is beyond the scope of this article, we would like to emphasize the importance of considering this crucial aspect, that is, feasibility. Training programs should take into account barriers such as faculty time constraints, residents' duty-hour regulations, and lack of familiarity of the faculty and residents with these tools. For successful implementation, program directors should select the tools that are both easy to use and for which there are well-established sources of validity evidence in the literature. Similar to the concept of evidence-based medicine, surgical educators, program directors, and other teaching faculty should embrace only these tools after significant and positive educational outcomes have been demonstrated. The aforementioned deliberations, in combination with faculty development, would most likely result in a high compliance rate in use of these tools by faculty.

ACKNOWLEDGMENTS

The authors acknowledge Foundation for Surgical Fellowships support for Dr. Ghaderi's fellowship training.

REFERENCES

1. Halsted WS. The training of the surgeon. In: Carmichael AG, Ratzan RM, eds. *Medicine: A Treasury of Art and Literature*. New York, NY: Harkavy Publishing Service; 1991:267–271.
2. Ulmer C, Wolman DM, Johns MM, eds. *Resident Duty Hours: Enhancing Sleep, Supervision, and Safety*. Washington, DC: The National Academies Press; 2009.
3. Vanderveen K, Bold RJ. Effect of generational composition on the surgical workforce. *Arch Surg*. 2008;143:224–226.
4. Troppmann KM, Palis BE, Goodnight JE, et al. Career and lifestyle satisfaction among surgeons: what really matters? The National Lifestyles in Surgery Today Survey. *J Am Coll Surg*. 2009;209:160–169.
5. Philibert I, Friedmann P, Williams WT. New requirements for resident duty hours. *JAMA*. 2002;288:1112–1114.
6. Ulmer C, Wolman DM, Johns MM, eds; Committee on Optimizing Graduate Medical Trainee (Resident) Hours and Work Schedule to Improve Patient Safety. *Resident Duty Hours*: *Enhancing Sleep, Supervision and Safety*. Washington, DC: The National Academies Press; 2008.
7. Lewis FR, Klingensmith ME. Issues in general surgery residency training. *Ann Surg*. 2012;256:553–559.
8. Nasca TJ, Philibert I, Brigham T, et al. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366:1051–1056.
9. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the 2010 Ottawa Conference. *Med Teach*. 2011;33:206–214.
10. Scott DJ, Dunnington GL. The new ACS/APDS Skills Curriculum: moving the learning curve out of the operating room. *J Gastrointest Surg*. 2008;12:213–221.

11. ACS/APDS Surgical Skills Curriculum for Residents. ACS Division of Education Web site. http://www.facs.org/education/surgicalskills.html. Updated August 16, 2012. Accessed October 1, 2013.
12. Kern DE, Thomas PA, Howard DM, et al. *Curriculum Development for Medical Education: A Six-Step Approach*. Baltimore, MD: Johns Hopkins University Press; 2009.
13. The American Board of Surgery, General Surgery Residents Assessment. American Board of Surgery Web site. http://www.absurgery.org/default.jsp?certgsqe_resassess. Accessed October 1, 2013.
14. Reznick RK. Teaching and testing technical skills. *Am J Surg*. 1993;165:358–361.
15. Fried GM, Feldman LS. Objective assessment of technical performance. *World J Surg*. 2008;32:156–160.
16. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, et al. Objective assessment of technical surgical skills. *Br J Surg*. 2010;97:972–987.
17. Ahmed K, Miskovic D, Darzi A, et al. Observational tools for assessment of procedural skills: a systematic review. *Am J Surg*. 2011;202:469–480.
18. Joint American Educational Research Association, *The Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
19. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull*. 1955;52:281–302.
20. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37:830–837.
21. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol*. 1995;50:741–749.
22. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119:166–167.
23. Schmitz CC. Your intergalactic decoder ring has arrived: reliability and validity defined. American College of Surgeons Residency Assist Page Web site. http://www.facs.org/education/rap/schmitz0506.html. Updated May 24, 2006. Accessed October 1, 2013.
24. Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. New York, NY: American Council on Education & Macmillan; 1989.
25. Kane MT. An argument-based approach to validation. *Psychol Bull*. 1992;112:527–535.
26. Kane MT. Validating interpretive arguments for licensure and certification examinations. *Eval Health Prof*. 1994;17:133–159.
27. Satava RM, Cuschieri A, Hamdorf J. Metrics for objective assessment of surgical skills workshop. Metrics for objective assessment. *Surg Endosc*. 2003;17:220–226.
28. Aggarwal R, Moorthy K, Darzi A. Laparoscopic skills training and assessment. *Br J Surg*. 2004;91:1549–1558.
29. Carter FJ, Schijven MP, Aggarwal R, et al; Work Group for Evaluation and Implementation of Simulators and Skills Training Programmes. Consensus guidelines for validation of virtual reality surgical simulators. *Surg Endosc*. 2005;19:1523–1532.
30. Korndorffer JR Jr, Kasten SJ, Downing SM. A call for the utilization of consensus standards in the surgical education literature. *Am J Surg*. 2010;199:99–104.
31. Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc*. 2003;17:1525–1529.
32. Vleuten CV. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996;1:41–67.
33. Cizek GJ, Rosenberg SL, Koons HH. Sources of validity evidence for educational and psychological tests. *Educ Psychol Meas*. 2008;68:397–412.
34. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20:1159–1164.
35. Datta V, Mackay S, Mandalia M, et al. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg*. 2001;193:479–485.
36. Francis NK, Hanna GB, Cuschieri A. The performance of master surgeons on the Advanced Dundee Endoscopic Psychomotor Tester: contrast validity study. *Arch Surg*. 2002;137:841–844.
37. Pellen MG, Horgan LF, Barton JR, et al. Construct validity of the ProMIS laparoscopic simulator. *Surg Endosc*. 2009;23:130–139.
38. Egi H, Okajima M, Yoshimitsu M, et al. Objective assessment of endoscopic surgical skills by analyzing direction-dependent dexterity using the Hiroshima University Endoscopic Surgical Assessment Device (HUESAD). *Surg Today*. 2008;38:705–710.
39. Chmarra MK, Klein S, de Winter JC, et al. Objective classification of residents based on their psychomotor laparoscopic skills. *Surg Endosc*. 2010;24:1031–1039.
40. Derossis AM, Fried GM, Abrahamowicz M, et al. Development of a model for training and evaluation of laparoscopic skills. *Am J Surg*. 1998;175:482–487.
41. Korndorffer JR Jr, Dunne JB, Sierra R, et al. Simulator training for laparoscopic suturing using performance goals translates to the OR. *J Am Coll Surg*. 2005;201:23–29.
42. Fried GM, Derossis AM, Bothwell J, et al. Comparison of laparoscopic performance in vivo with performance measured in laparoscopic simulator. *Surg Endosc*. 1999;13:1077–1081.
43. Sroka G, Feldman LS, Vassiliou MC, et al. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. *Am J Surg*. 2010;199:115–120.
44. Ghaderi I, Vaillancourt M, Sroka G, et al. Performance of simulated laparoscopic incisional hernia repair correlates with operating room performance. *Am J Surg*. 2009;201:40–45.
45. Williams RG, Verhulst S, Colliver JA, et al. A template for reliable assessment of resident operative performance: assessment intervals, numbers of cases and raters. *Surgery*. 2012;152:517–527.
46. Sedlack RE. The Mayo Colonoscopy Skills Assessment Tool: validation of a unique instrument to assess colonoscopy skills in trainees. *Gastrointest Endosc*. 2010;72:1125–1133.
47. Gofton WT, Dudek NL, Wood TJ, et al. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. 2012;87:1401–1407.
48. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.
49. Downing SM, Yudkowsky R. *Assessment in Health Professions Education*. New York, NY: Routledge; 2009.
50. Van Sickle KR, Baghai M, Huang IP, et al. Construct validity of an objective assessment method for laparoscopic intracorporeal suturing and knot tying. *Am J Surg*. 2008;196:74–80.
51. Palter VN, MacRae HM, Grantcharov TP. Development of an objective evaluation tool to assess technical skill in laparoscopic colorectal surgery: a Delphi methodology. *Am J Surg*. 2011;201:251–259.
52. Vassiliou MC, Kaneva PA, Poulose BK, et al. Global Assessment of Gastrointestinal Endoscopic Skills (GAGES): a valid measurement tool for technical skills in flexible endoscopy. *Surg Endosc*. 2010;24:1834–1841.
53. Ghaderi I, Vaillancourt M, Sroka G, et al. Evaluation of surgical performance during laparoscopic incisional hernia repair: a multicenter study. *Surg Endosc*. 2011;25:2555–2563.
54. Zeller RA, Carmines EG. *Measurement in the Social Sciences: The Link Between Theory and Data*. Cambridge: Cambridge University Press; 1980.
55. Kurashima Y, Feldman LS, Al-Sabah S, et al. A tool for training and evaluation of laparoscopic inguinal hernia repair: the Global Operative Assessment of Laparoscopic Skills-Groin Hernia (GOALS-GH). *Am J Surg*. 2007;201:54–61.
56. Moorthy K, Munz Y, Dosis A, et al. Bimodal assessment of laparoscopic suturing skills. *Surg Endosc*. 2004;18:1608–1612.
57. Nichols PD, Williams N. Consequences of test score use as validity evidence: roles and responsibilities. *Educ Meas*. 2009;28:3–9.
58. Mehrens WA. The consequences of consequential validity. *Educ Meas*. 2005;16:16–18.
59. Shepard LA. The centrality of test use and consequences for test validity. *Educ Meas*. 2005;16:5–24.
60. Chipman JG, Schmitz CC. Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. *J Am Coll Surg*. 2009;209:364–370.
61. Scott DJ, Goova MT, Tesfay ST. A cost-effective proficiency-based knot-tying and suturing curriculum for residency programs. *J Surg Res*. 2007;141:7–15.
62. Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. *Am J Obstet Gynecol*. 2006;195:617–621.
63. Shippey S, Handa VL, Chen TL, et al. Validation of an instrument for evaluation of subcuticular suturing using a plastic tissue model. *J Surg Educ*. 2009;66:31–34.
64. O'Connor HM, McGraw RC. Clinical skills training: developing objective assessment instruments. *Med Educ*. 1997;31:359–363.
65. Anastakis DJ, Wanzel KR, Brown MH, et al. Evaluating the effectiveness of a 2-year curriculum in a surgical skills center. *Am J Surg*. 2003;185:378–385.

66. Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. *Am J Surg*. 2007;193:551–555.
67. Wilasrusmee C, Lertsithichai P, Kittur DS. Vascular anastomosis model: relation between competency in a laboratory-based model and surgical competency. *Eur J Vasc Endovasc Surg*. 2007;34:405–410.
68. Wilasrusmee C, Phromsopha N, Lertsitichai P, et al. A new vascular anastomosis model: relation between outcome and experience. *Eur J Vasc Endovasc Surg*. 2007;33:208–213.
69. Leong JJ, Leff DR, Das A, et al. Validation of orthopaedic bench models for trauma surgery. *J Bone Joint Surg Br*. 2008;90:958–965.
70. Cass OW, Freeman ML, Peine CJ, et al. Objective evaluation of endoscopy skills during training. *Ann Intern Med*. 1993;118:40–44.
71. Chak A, Cooper GS, Blades EW, et al. Prospective assessment of colonoscopic intubation skills in trainees. *Gastrointest Endosc*. 1996;44:54–57.
72. Datta V, Bann S, Aggarwal R, et al. Technical skills examination for general surgical trainees. *Br J Surg*. 2006;93:1139–1146.
73. Shah J, Munz Y, Manson J, et al. Objective assessment of small bowel anastomosis skill in trainee general surgeons and urologists. *World J Surg*. 2006;30:248–251.
74. Bann S, Kwok KF, Lo CY, et al. Objective assessment of technical skills of surgical trainees in Hong Kong. *Br J Surg*. 2003;90:1294–1299.
75. Munz Y, Moorthy K, Bann S, et al. Ceiling effect in technical skills of surgical residents. *Am J Surg*. 2004;188:294–300.
76. Vick LR, Vick KD, Borman KR, et al. Face, content, and construct validities of inanimate intestinal anastomoses simulation. *J Surg Educ*. 2007;64:365–368.
77. Sidhu RS, Vikis E, Cheifetz R, et al. Self-assessment during a 2-day laparoscopic colectomy course: can surgeons judge how well they are learning new skills? *Am J Surg*. 2006;191:677–681.
78. Palter VN, Grantcharov TP. A prospective study demonstrating the reliability and validity of two procedure-specific evaluation tools to assess operative competence in laparoscopic colorectal surgery. *Surg Endosc*. 2012;26:2489–2503.
79. Sarker SK, Kumar I, Delaney C. Assessing operative performance in advanced laparoscopic colorectal surgery. *World J Surg*. 2010;34:1594–1603.
80. Miskovic D, Wyles SM, Carter F, et al. Development, validation and implementation of a monitoring tool for training in laparoscopic colorectal surgery in the English National Training Program. *Surg Endosc*. 2011;25:1136–1142.
81. Santos BF, Reif TJ, Soper NJ, et al. Development and evaluation of a laparoscopic common bile duct exploration simulator and procedural rating scale. *Surg Endosc*. 2012;26:2403–2415.
82. Szalay D, MacRae H, Regehr G, et al. Using operative outcome to assess technical skill. *Am J Surg*. 2000;180:234–237.
83. Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg*. 2007;204:308–313.
84. Dath D, Regehr G, Birch D, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc*. 2004;18:1800–1804.
85. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190:107–113.
86. Leblanc F, Zeinali F, Marks J, et al. Stepwise Assessment Tool of Operative Skills (SATOS): validity testing on a porcine training model of open gastrectomy. *J Am Coll Surg*. 2010;211:672–676.