Implementing Assessment Methods in Plastic Surgery

Amanda Gosman, M.D.
Karen Mann, M.Sc., Ph.D.
Christopher M. Reid, M.D.
Nicholas B. Vedder, M.D.
Jeffrey E. Janis, M.D.

San Diego, Calif.; Halifax, Nova Scotia, Canada; Seattle, Wash.; and Columbus, Ohio

Background: Principles of effective assessment have become increasingly popular topics in graduate medical education. Changes in the structure of plastic surgery training demand a thorough understanding of the state-of-the-art in assessing surgical trainees. Moreover, the authors’ understanding of different domains and methods of assessment and the available tools continues to grow.

Methods: The authors reviewed the available literature regarding assessment in graduate medical education, specifically as it pertains to plastic surgery. In addition, the authors present principles of effective assessment and report on the currently available assessment methods.

Results: Assessment is multifaceted and impacts everyone, not just the individual learner. For assessments to be useful, they need to possess validity and reliability. Moreover, there is a necessary pragmatism limiting different methods and tools for assessing learners. Some types of assessment are universally familiar and include examples such as written examinations and procedural logs. Other emerging areas that are actively being researched involve simulation, nontechnical skills, and procedure-specific technical assessments.

Conclusion: Updating the thoroughness and multidimensionality with which plastic surgery trainees are assessed is an evolving area and one that is ripe for continued research. (Plast. Reconstr. Surg. 137: 617e, 2016.)

From the Division of Plastic Surgery, University of California, San Diego; the Division of Medical Education, Faculty of Medicine, Dalhousie University; the Division of Plastic Surgery, University of Washington; and the Department of Plastic Surgery, The Ohio State University Medical Center.

Received for publication July 7, 2015; accepted November 9, 2015.
Copyright © 2016 by the American Society of Plastic Surgeons
DOI: 10.1097/01.prs.0000479968.76438.27

Disclosure: The authors have no pertinent financial disclosures.
prior research in medical education. However, over time, these measures have been deemed inadequate to fully assess graduating physicians.5,6 Our ability to fully train new physicians has been impacted by work-hour restrictions, funding shortages, and increased breadth of practice. Increased public pressure regarding outcomes and more scrutiny in supervision has also influenced training models to ensure that graduates were competent for individual practice. This resulted in a 2009 release by the Accreditation Council for Graduate Medical Education of the Next Accreditation System,7 that introduced the concept of educational “milestones” and constituted a new model of competency-based assessment. Milestones are developmentally based markers that represent proficiency in specialty-specific competencies that the trainee has achieved at established intervals as they progress through training.7 The intent is to have a dynamic model to track progression through training, with the ability to accurately identify inadequacies early and ensure appropriateness for promotion. The model is fundamentally centered on competency rather than time-based training. This new system has been implemented in seven different specialties, and recently initial results of validity have been reported.8

The work of the Accreditation Council for Graduate Medical Education is necessary and has been performed in good stewardship of graduate medical education. However, their efforts are broad and aimed to include all medical specialties, the majority of which are nonsurgical. Unfortunately, there is a lack of specific methods for assessment that apply to the training of surgeons. It is widely accepted that many surgical graduates are not prepared to practice independently.5 As a result, much research has emerged focusing on both technical skill, particularly relevant in surgical education, and nontechnical skills. Many different assessment tools are being developed to evaluate these domains and, currently, this is a rich area of research. In this article, we present a background on the taxonomy and nomenclature used in assessing surgical trainees, consider current principles of assessment, and review some validated assessment tools currently in use.

### APPRAOCHES TO ASSESSMENT

In 1990, Miller described his oft-cited hierarchical four-level pyramid (Fig. 1) describing progression of competence where one “knows,” then “knows how,” then “shows how,” and finally “does.”9 This is the intuitive structure of development that many practicing surgeons experienced as they progressed during training. However, over time, investigators have better begun to describe the importance of using formative and summative evaluations,4 the differences and importance of both technical10–12 and nontechnical13–15 skills, and the differing assessment needs at differing levels of training.16

In 2010, attendees at the Ottawa Conference developed a consensus statement and recommendations17 that provided a set of overarching principles for good assessment. These criteria address examinees, patients, teachers, educational institutions, health care systems, and regulators. It is clear that assessment cannot be thought of as measurement alone. It has important roles in guiding learning, in assessing worth or value, and in improving quality (Fig. 2).

Van der Vleuten and Schuwirth described a utility of assessment equation18 that was later modified by Swing et al.19 and includes reliability, validity, ease of use, resources required, ease of interpretation, and educational impact. We describe each of these briefly below.

### Reliability

Reliability is the extent to which an assessment can be reproduced, if administered again, or by another observer (intrarater reliability).20 For example, will another observer watching the same resident performing an operation make the same
Reliability is also a function of how well the domain being tested is being sampled. For example, a multiple-choice question examination limited to 30 questions, or an objective structured clinical examination of only three stations, may not allow for adequate sampling of the outcomes the learner is expected to achieve. A third aspect of reliability is the extent to which we can generalize from the scores we have to the entire domain of competence. When our assessment samples from only a limited area, can we be sure that we have observed enough instances of the individual learner’s performance to feel confident that they are also competent at tasks that were not observed? For example, evaluating proficiency at vascular anastomosis may not carry over to aesthetic expertise.

Validity

Validity refers to the extent to which there is certainty that the inferences we are making from assessment are justified. Assessing validity is about gathering evidence to support our judgments and is a function of using the appropriate assessment approach and drawing appropriate conclusions. A common misconception is that some assessment tools are more valid than others: in fact, there are different facets of validity as described by Moorthy et al.16 Construct validity is the extent to which a test measures its intended target. Content validity is the extent to which the domain that is being measured is truly being measured by the tool and not by means of a surrogate marker. Face validity is the extent to which the examination resembles real-life situations. Predictive validity is the ability of the assessment to correlate with future performance. Criterion validity refers to the relationship of the test scores to an established “gold standard.” Newer thinking about validity includes all of these aspects of validity as part of construct validity.17 Rather than different types of validity, they are considered as different processes of validation.

Ease of Use

Ease of use of assessment must consider the learner, the assessor, and the system as a whole. Tools developed must be easily implemented and manageable by evaluators and trainees.

Resources Required

Resources required and a method’s cost effectiveness are particularly relevant in the current graduate medical education environment, as it is often the limiting step. Resources are both human and material, and include such considerations as faculty time and training required.

Ease of Interpretation

Ease of interpretation refers to the extent to which the test provides information that is meaningful and that is presented in such a way that both faculty and learners may understand it. Meaningful data require attention to what information is collected and its clear connection to the program’s objectives.
Educational Impact

Educational impact of assessment describes its impact on learning both before and after the assessment; this impact should be positive and reciprocal. Questions to consider include how the assessment influences what and how learners prepare, and how it influences their ability to learn from it and improve. Assessors should also receive feedback and education regarding how to effectively evaluate, and educational systems can be improved and refined when data and feedback are collected.

PRINCIPLES OF ASSESSMENT

Assessments can be characterized as formative or summative. The goal of formative assessment is to provide feedback with the intent to guide or inform the learner about how to improve his or her performance. In contrast, the purpose of summative assessment is to form a judgment of a learner’s performance and whether that performance has met the required criteria. Generally, any assessment method can be used for both purposes. In fact, good educational practice would suggest that formative assessments should mirror summative assessments, both to give the learner experience and some guidance as to how the summative assessment will occur.

Previous authors have proposed principles of assessment in medical education that affects different participants in the process with broad impacts. When considering the trainee, the goals are to provide feedback, direction, and opportunities for learning. Considered together, the assessment should foster the trainee’s progression. In addition, assessments can inform the curriculum by identifying weakness, creating coherence, allowing for cross-validation, and establishing standards and opportunity for refinement. When considered at the institutional level, assessments allow for institutional self-reflection, discrimination among candidates, developing and sharing educational values, providing data for research, and promoting faculty development. For the public, assessments allow for certification of competence.

TYPES OF ASSESSMENT

There are many different means of performing an assessment that range from traditionally accepted practices to modern techniques derived from scientific inquiry. In this section, we review and describe currently used assessment measures.

Written Examinations

Written examinations to test learners’ knowledge, synthesis, and judgment generally follow well-established and accepted evidence. The most familiar example would be multiple-choice question examinations. Annual written in-service examinations developed by the American Board of Surgery and the American Society of Plastic Surgeons are being continually enhanced and have demonstrated content and construct validity. However, it has been shown that the written examinations do not correlate with technical skill, a central tenet of surgical competence.

Procedural Logs

Another measure in use is review of procedural logs, which have not been shown to correlate with competence. Requirements for attaining certain arbitrarily picked procedural numbers say nothing about one’s aptitude or performance. Moreover, studies have shown that graduates are shockingly inexperienced in procedures deemed necessary by practicing surgeons, and that there is a continued downward trend in case numbers by postgraduate year–5 general surgery residents and first assistant cases by postgraduate year–1 residents. This same situation has been observed in plastic surgery trainees.

Core Competencies

Other measures currently in use are subjective in nature; the most accepted measure is assessment of trainees in each of the Accreditation Council for Graduate Medical Education six core competencies (Table 1). Although these are ubiquitous and have been a mandate of the graduate medical education governing bodies, there is no specific requirement that ensures appropriate training or uniform implementation by evaluators and programs. Historically, these core competencies include a subjective global assessment of trainees’ fitness for promotion. This subjective assessment lacks any validated structure or required components, and is subject to multiple biases. There are no specific requirements for amount or type of observation that assessors must achieve to make a global assessment. Considered together, these factors provided the impetus for the development of the Next Accreditation System.

Direct Observation and Task/Domain-Specific Assessment

An essential, historical part of surgical training is the direct observation and evaluation of
trainees, primarily in the operating room. A familiar phrase, “see one, do one, teach one,” has been the slogan for decades for both practicing and retired surgeons. However, as stated earlier, trainees are performing fewer cases, likely because of decreased time in training and the growing number of cases performed within each specialty diluting out the number of index cases.

Another widely accepted method of direct observation, used by the sponsors of the United States Medical Licensing Examination, is the Objective Structured Clinical Examination.28 The Objective Structured Clinical Examination is used commonly by medical schools across disciplines and is frequently a mandatory part of the curriculum, although currently there is no uniform or validated implementation of the Objective Structured Clinical Examination for surgical trainees. Models have been proposed in general surgery29 and plastic surgery.30

The observation and feedback that occurs during the course of work is described as work-based assessment.31 This type of assessment offers an opportunity to assess trainees’ ability to meet required milestones in the course of their authentic work. Kogan and colleagues conducted a systematic review of the strengths and weaknesses of available tools for direct observation, finding scarcity in validity and published educational outcomes.32 Another recent review regarding direct observation and assessment of psychomotor skills used a grading system for assessment tools, developed by Swing et al. and adopted by the Accreditation Council for Graduate Medical Education.19 Of the 30 studies meeting their inclusion criteria of over 1400 cited in the literature, only two received a class 1 Accreditation Council for Graduate Medical Education grade and are recommended based on level A Accreditation Council for Graduate Medical Education evidence. These were the seven-item Global Rating Scale and the Procedure-Based Assessment.33 Both metrics are well structured and defined to allow for specific and focused evaluation of the trainee in certain situations, real or simulated. They have been adapted to many different skills and specialties.

Evidence-based assessment of technical skill, particularly relevant to plastic surgery, has been performed with the use of the Objective Structured Assessment of Technical Skills.12 The Objective Structured Assessment of Technical Skills was originally developed to broadly assess surgical performance at the bench side and over time its framework has been modified to a vast array of surgical procedures in real scenarios. This tool is being refined for use with specialty-specific development for index procedures. However, its application is limited, because of significant costs associated with its use. Expenses monetarily and in time and resources make it impractical for many programs to implement in evaluations. However with time, continued refinement at advanced centers, and a settling in of competency-based evaluations, it will likely become more common.

Much work has also gone into developing tools to evaluate nontechnical skills such as professionalism, communication, teamwork, leadership, and decision-making. These qualities are particularly important for future successful practice of trainees and are often overlooked in assessment. Two of the most recognized tools that have been developed and investigated are the Observational Teamwork Assessment for Surgery34,35 and Non-Technical Skills for Surgeons.13 However, these tools are still being studied and refined to ensure validity in the surgical realm. To date, there have not been any plastic surgery–specific metrics to assess technical or nontechnical skill, although technical skills have been addressed in the context of milestones in the Next Accreditation System.

**SIMULATION**

Simulation has been growing as an effective means of learning and teaching through structured, repetitive, safe practice.36–38 It has been shown to be an effective and good method for assessing not only technical skills but also judgment, teamwork, and cognitive skills. In addition, the safe practice it allows has been shown to transfer to real-world practice and improve outcomes.39,40 However, expenses relating to cost and time limit its feasibility. Moreover, there has yet to be a uniform, structured simulation tool that has been proven reliable and valid in a variety of contexts. Sadly, recent surveys of surgical programs have shown that a minority have or use their simulation centers, and of those that have them available, few have any established curricula.41,42 Nevertheless, in one survey, 100 percent of program directors were interested in a unified, structured curriculum.41

**CONCLUSIONS**

The shift toward competency-based assessment has taken place. Accompanying this have
been tremendous efforts to define, refine, and develop assessment tools that are both reliable and valid. Currently, the work continues, with certain methods continuing to rise above the rest. Regardless, it appears that as long as we ensure multiple observations, and use multiple methods of assessment, we can possibly compensate for the individual flaws of each tool.2,44

Amanda Gosman, M.D.
Rady Children’s Hospital San Diego
200 West Arbor Drive
San Diego, Calif. 92103-8890
agosman@ucsd.edu

REFERENCES


