A Unified Approach to Validation, Reliability, and Education Study Design for Surgical Technical Skills Training

Robert M. Sweet, MD; David Hananel, BS; Frances Lawrenz, PhD

Objective: To present modern educational psychology theory and apply these concepts to validity and reliability of surgical skills training and assessment.

Design: In a series of cross-disciplinary meetings, we applied a unified approach of behavioral science principles and theory to medical technical skills education given the recent advances in the theories in the field of behavioral psychology and statistics.

Conclusions: While validation of the individual simulation tools is important, it is only one piece of a multimodal curriculum that in and of itself deserves examination and study. We propose concurrent validation throughout the design of simulation-based curriculum rather than once it is complete. We embrace the concept that validity and curriculum development are interdependent, ongoing processes that are never truly complete. Individual predictive, construct, content, and face validity aspects should not be considered separately but as interdependent and complementary toward an end application. Such an approach could help guide our acceptance and appropriate application of these exciting new training and assessment tools for technical skills training in medicine.

Arch Surg. 2010;145(2):197-201

In this article, we focus on curricular design and methodologies that examine validity and reliability of technical skills training and assessment. The assessment of health care professionals is part of the training process and presents as a continuum of benchmarks throughout an individual’s career (Figure). In fact, if we consider both formative and summative assessment, training and assessment become almost indistinguishable from each other and combine with the daily performance of the individual.

See Invited Critique at end of article

Technical skills performance and assessment can be broken down into cognitive, psychomotor, communication, and affective domains, as implicitly reflected in the 6 core competencies defined by the Accreditation Council for Graduate Medical Education.1 With written and oral examinations as its cornerstone assessment tool, a taxonomy for the cognitive domain was established as early as 1956,2 whereas a taxonomy for the psychomotor domain did not emerge until 1966.3 In the surgical disciplines, the emergence of early computer-based simulation tools in 1993 and the work of Richard Reznick, MD, in Toronto, Canada, in 1997 enabled the consideration of objective assessment of technical skills.5,3 Since then, much effort has been focused on validating the individual simulator that had been developed to train and assess technical skills. Most simulator development was driven at first by the demonstration of engineering principles (early 1990s) and subsequently by achieving clinical reality (late 1990s to early 2005). We are now appropriately transitioning toward a more educationally minded design with virtual mentorship and instruction.

To date, validity has been claimed by many trainers and this catch-all term has led to justification for use across different points of the medical education continuum, as long as a semblance of face, content, construct, and predictive validity aspects had been shown. The problem with this approach is that questions of assessment used at various benchmarking events during the career of a health care professional are distinct while validity is cumulative. There is a different level of burden of proof for high- and low-stakes assessments along the continuum (Figure).
If we consider the case of the Fundamentals of Laparoscopic Surgery course developed by the Society of American Gastrointestinal and Endoscopic Surgeons, we get a glimpse of the thought process and effort it takes to validate even a simple test to be used as a high-stakes examination by a professional society.

Gallagher et al have established an important foundation for understanding these concepts. They describe discrete benchmarks to guide studies, described as face validity, content validity, construct validity, concurrent validity, discriminate validity, and predictive validity. This seminal work guided almost all validity studies in surgical technical skills training to date and is based on validation criteria established by the standards set by the American Psychological Association in 1974. Most of these studies focus on validating individual simulators, not curricula. Currently, the new American Psychological Association standards describe validity as a continuing argument, not one benchmarked as above, but one that provides evidence to support the notion that the process, instrument, or device is indeed assessing what it is purported to measure. As stated in the Standards for Educational and Psychological Testing, “The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations.” At its root, validation is determined by evidence.

Construct validity, therefore, is the totality of validity and as such, has 6 aspects: (1) content (how adequately the items on an assessment represent the domain of the construct being measured), (2) substantive (to what extent the assessment items reflect an individual’s actual performance), (3) structural (how appropriately the scores or ratings represent the domain of the construct being measured), (4) generalizability (how similarly individuals or groups perform on the assessment and how appropriate the assessment is for all individuals in all settings), (5) external (how well the items on an assessment compare with the items on similar instruments), and (6) consequential (if the items on the assessment are fair, unbiased, and useful and what the intended or unintended impact on the student, faculty, patients, or society is). These 6 aspects therefore must be viewed as interdependent and complementary forms of validity evidence and not as separate and substitutable validity types.

In line with Downing’s description of Messick’s theory for medical education, the objective of this article is to present modern educational psychology theories and apply these concepts to the integration of simulation education curricular design with validity and reliability of surgical skills training and assessment.

METHODS

In a series of 6 monthly cross-disciplinary meetings, experts representing the area of surgical education, simulation design, and educational psychology, we expanded on the foundation laid out by Wiggins and McTighe as well as Gallagher et al and apply a unified approach of behavioral science principles and theory to surgical skills education. This theory is primarily based on Messick’s unified concept of validity: validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. When applying this unified concept, we should not focus on designing studies to establish various types of validity, but rather, we should design and study a series of research questions that best satisfy the end goal of the designed evaluation tool.

RESULTS

If we approach the task from an educational perspective, we need to consider the overall design, with validity studies running concurrent with design development. According to the backward design principle of Wiggins and McTighe, we must first determine the purpose and desired learning outcome and work backward by learning objectives targeted toward accomplishing these outcomes, considering that there are multiple potential uses for simulation training tools along the timeline of surgical education (Table). The desired outcome drives the development of learning objectives that are developed, as described by Prideaux. For technical skills training, objectives can be derived via answering the following questions: What does it mean to be safe and efficient? What does it mean to perform a procedure independently? Which learning domains predominates (eg, cognitive, psychomotor, or communication)?

These first 2 steps in turn guide the study design or how it is that the subject matter should be introduced and in what format. Much like clinical assessment studies, the validity and reliability of educational research findings depend on 2 issues: the research design and the methods of assessment. Both are necessary to produce high-quality information. Research design assures the optimal arrange-
A wide variety of research designs are possible depending on the types of questions being addressed, e.g., experimental designs, quasi-experimental designs, multiple baseline designs, treatment reversal designs, interrupted time series designs, correlational, or predictive designs as well as qualitative approaches. Designs should be shaped based on the context and the intended sample (e.g., health professional students vs residents vs practitioners) and include enough sub-study samples for making inferences, while methods of assessment assure the legitimacy of the data.

In a case of a simulation exercise, a variety of answer choices were chosen by at least some of the respondents. In the case of a simulation exercise, this analysis would show that all of the differences in question is whether or not the response options function as test-retest. The test-retest method was really designed to assess the stability of responses over time. In other words, if a test taker responded one way today would she respond that way again on a different occasion?

Another element would involve defining the standards. Criteria clearly need to be established to determine relative competence and mastery of skills. This can be established initially with a panel of experts, but ultimately should be determined and benchmarked with learning curves established clinically in the literature and by data generated by the simulation curricula over time.

Much as the standards will evolve over time, so will the curriculum. Each version can and should reflect such changes as well as the unique regional populations and requirements of its learners, and these changes will need to be examined for validity. Techniques and practice patterns vary from institution to institution, and what may be valid for a population of surgical residents in Brazil may have little relevance for practicing surgeons in China, though the basic principles may remain the same. These issues provide further credence to the principle that continued evidence must be collected to maintain validity.

Another element involves collection of pilot data and the establishment of consensus and reliability between evaluators. Once the curriculum is in the best possible format, pilot data should be collected from people similar to the intended sample (health professional students vs residents vs practitioners) and include enough subjects to determine how well the items function. As mentioned before, this will allow for the establishment of preliminary standards, and subject matter experts can help to refine the curriculum. The pilot data can be used to answer a variety of questions about the trainer. One critical question is whether or not the response options function adequately. In other words, do they represent the range in the sample? If the instrument were a multiple choice test, this analysis would show that all of the different answer choices were chosen by at least some of the respondents. In the case of a simulation exercise, common error-prone paths will be chosen by less experienced or less skilled practitioners, hopefully at a rate that is similar to that in the real environment. A carefully designed simulator with metrics derived from the real environment can help get simulation developers “in the ballpark” and the parameters can be fine-tuned once the preliminary data are analyzed to represent a realistic sample. Once repeated, the distribution of errors and correct maneuvers in the trainer should more appropriately represent what occurs in the real environment.

An assessment device cannot be valid if it is not reliable. Reliability is a necessary but insufficient condition for validity. Gallagher et al. suggest 2 common methods to establish intrarater reliability—the split-half test and test-retest. The test-retest method was really designed to assess the stability of responses over time. In other words, if a test taker responded one way today would she respond that way again on a different occasion?

The split-half test and α assess the consistency of the responses to items on the device with each other. For example, are an individual person’s responses to the set of items similar? Consistency within an instrument is important if the instrument is supposed to be measuring a single concept. If the instrument is designed to measure different concepts, such as a survey with multiple parts, an overall α measure is not appropriate. However, it might be appropriate to have reliability measures for parts of a survey.

Meetings among key technical thought leaders initially helped to establish the conditions, metrics, and standards, and by doing this early, it becomes easier to establish interrater reliability. This is especially important when using subjective evaluations as part of the curriculum, as there are multiple correct ways and techniques to perform a procedure. For example, one surgeon might provide a top score to a student for a performance that mimicked his or her own technique, while giving a poor grade to another candidate who performs a procedure using a different, albeit established, technique. The earlier questions—about what it means to be safe and efficient and to perform a procedure independently—should guide such evaluations instead of surgeons asking themselves whether or not they would perform a surgery the same way as their students. Interrater reliability is then examined by looking for agreement among different evaluators for each item for a variety of cases. $R > .8$ is a common standard.

Acceptability is extremely important, but does not contribute to validity. Often the first thing we do when we are considering a new curriculum is establish acceptability, and while this is not scientifically valid, it can be useful on a surface level. Some examples of common acceptability questions are whether the tool is easy to use and implement and whether others are using it. Other examples include questioning whether the tool is more enjoyable, practical, accessible, and/or affordable than current methods of training. Given the paucity of validation data, it is interesting that these questions are currently guiding our purchases or use of commercially available simulation tools in our new curricula.

A practical example of a simulation-based central line placement curriculum can demonstrate these concepts.
In this example, our intent was to see if it was appropriate for the trainer to serve as an assessment/training tool during residency clinical skills curricula. Because validation is a continuous process that is never complete, a program asks itself what conditions need to be satisfied to make them feel comfortable using this tool for this purpose. Is the curriculum’s educational content appropriate and up to date? Are the techniques being taught properly and are they at least the standard of care? These research questions can be addressed by running content matter experts through the simulation-embedded curriculum as illustrated above and soliciting their input. Central (Accreditation Council for Graduate Medical Education) vs local (individual program) consensus and direction is ideal; the relative role of each is beyond the scope of this article. Are the methods that are being used to train students in a technique sound and reproducible in a real environment? This question can be addressed through demonstration of the translation of skills from the simulation laboratory to the clinical environment, much as it was presented by Seymour et al.21 Can the program shorten the learning curve compared with standard training techniques? One approach addressing this question is a longitudinal design in which residents are randomized to training vs no training and clinical performance is compared at multiple times. Training-transfer ratios may be established by training residents to different benchmarks on the simulator and then comparing this with results in the clinical environment. A calculation of the amount of time spent in the simulated environment with that in the real environment can be made. Prior to purchase, acceptability questions will invariably be addressed as well, such as whether the training program is easy to use and whether it is being used and by whom. Cost-benefit ratios are compared with other modalities of training.

Once these studies are completed, there may or may not be strong evidence for integration into a residency clinical skills curriculum. There may even be preliminary evidence for using the program for credentialing, especially if learning curves and standard deviations have been established, but this level of validity has not completely been satisfied. There are different research questions that may better address this purpose.

With direction of the appropriate hospital vs governing medical boards, some additional questions that would need to be answered to consider such a trainer for credentialing may include the following: Does it predict performance in the real environment? More specifically, are errors that are consistently being made on real cases also occurring on the training model? This can be answered by setting up specific scenarios in the simulated environment and correlating metrics in the clinical environment. Is it a useful addition to the current methods of credentialing (ie, number of supervised cases performed, standardized boards examinations, and recertification examinations)? Can error reduction, efficiency, and patient safety outcome improvements be demonstrated through participation in the program? This would be determined by examining intermediate clinical metrics as well as patient outcomes data among practitioners using simulation training strategies vs those who do not. Correlations with board scores and other measurement benchmarks could be examined. If such questions were addressed and shown to be true, there is strong evidence for implementation of a training curriculum for credentialing purposes.

Considering the use of such a curriculum for selecting residents may require that additional challenges to validity be satisfied. Can the program predict performance during and at the conclusion of formal training? Are the skill sets that are being measured comprehensive and critical to success as a practitioner? Can the program detect and dissect skills that are trainable vs those that are inherent and untrainable? This would require the implementation of a simulation tool for a long time and understanding how it relates to a real learning curve and appropriate standard deviations along that learning curve to truly be able to determine if a single exercise or brief series of exercises could predict future core abilities. We would also argue that consideration for such a use should require a higher level of reliability.

While validation of the individual simulation tools is important, it is only one piece of the overall multimodal curriculum that deserves examination and study. We propose concurrent validation throughout the design of simulation-based curriculum rather than once it is complete. We described an application of Wiggens and McTigue’s backward design approach to curriculum development for technical skills starting with outcomes and tailoring research questions as burden of proof toward this purpose. We also build on Gallagher’s foundation of validity of simulation tools by introducing Messick’s modern unified concepts to this burgeoning field and have provided some core general research approaches and questions that could be applied to questions of validity with procedural simulators. We embrace the concept that validity and curriculum development are interdependent, ongoing processes that are never truly complete. Individual predictive, construct, content, and face validity aspects should not be considered separately but together as interdependent and complementary toward an end application. Such an approach could help guide our acceptance and appropriate application of these exciting new training and assessment tools for technical skills training in medicine.

Accepted for Publication: February 10, 2009.
Correspondence: Robert M. Sweet, MD, Department of Urologic Surgery, University of Minnesota, Mayo Mail Code 394, 420 Delaware St SE, Minneapolis, MN 55455 (rsweet@umn.edu).

Author Contributions: Study concept and design: Sweet, Hananel, and Lawrenz. Acquisition of data: Sweet. Analysis and interpretation of data: Sweet. Drafting of the manuscript: Sweet, Hananel, and Lawrenz. Critical revision of the manuscript for important intellectual content: Lawrenz. Statistical analysis: Lawrenz. Administrative, technical, and material support: Sweet and Hananel. Study supervision: Sweet.
Validity in Educational Research

Critically Important but Frequently Misunderstood

This article addresses 2 important issues in surgical education: (1) how we design educational curricula and (2) how we assess validity of educational assessments. In the past, many training programs purchased simulators without clear educational goals or objectives. In research terms, this is equivalent to starting a research program without a hypothesis or study methodology, a situation guaranteed to lead to poor results. More recently, surgical educators have begun to focus on educational curricula rather than on simulators. Curriculum design for technical skill education involves setting goals and objectives at the start, designing interventions targeted to those goals, and developing assessment tools that can certify competency in the desired skills. Depending on the goals, curricula may or may not involve the use of simulators. The authors of this article call this process a backward design, because they start with the end goal (a well-trained surgeon) and work backward to develop curricula to meet that goal. Rather than being backward, this could more properly be considered forward thinking.

In the world of surgical education, validity is often discussed but frequently misunderstood. We often call a simulator or test valid on the basis of 1 study, in 1 population, and with testing of only 1 kind of validity. Instead, validity should be considered similar to the levels of evidence that we apply to clinical research. As the authors of this study point out, validity must be based on multiple sources of evidence and may only apply in certain populations or circumstances. Additionally, validity alone is not sufficient. In clinical studies we sometimes say that a result is statistically significant but not clinically relevant. An educational assessment may similarly be valid but unreliable or irrelevant. Again, this points out the critical importance of building educational interventions around clear goals and objectives.

Andrew S. Wright, MD

Author Affiliation: Department of Surgery, University of Washington, Seattle.
Correspondence: Dr Wright, Department of Surgery, University of Washington, 1595 NE Pacific St, Box 35410, Seattle, WA 98195 (awright2@u.washington.edu).